NYBG NEW YORK BOTANICAL GARDEN

Herbarium Collections Data Management

Kimberly Watson, Assistant Director of Biodiversity Information Management
Joel Ramirez, Web Developer for Biodiversity Information Management
NYBG Herbarium Training Workshop
11 June 2025, 09 July 2025

Overview

Collections Information

Collections Management Systems (CMS)

Biodiversity Data Standards

Authority Files and Controlled Vocabularies

Data Cleanup

Data Aggregators

Data Mobilization

Collections Information

Collections Information: Value in Digitization

Information important for organizing, preserving, utilizing the collections

- **Record keeping and object tracking**
  - Detailed documentation of each object
  - Collection inventory
  - Activities, processes, workflows
- **Access and Use**
  - Facilitate research
  - Facilitate data sharing
  - Elevate museum/collection profile, public awareness

Nelson Gil and Ellis Shari. 2019 The history and impact of digitization and digital data mobilization on biodiversity research. *Phil. Trans. R. Soc. B* 374: 20170391 http://doi.org/10.1098/rstb.2017.0391

NYBG

Types of Collections Information

- **Occurrence data:** evidence of an occurrence of a species by a person (or instrument) at a particular place and time.
  - Object type, Catalog Number, Accession Number
  - Collector(s), Number, Date, Location, Georeference
  - Taxon, Identified By, Date, Determination History
- **Digital Multimedia resources**
- **Physical location, filing system**
- **Processing, transactions, movements (internal/external)**
- **Rights, permissions, provenance, permits**
- **Condition checks, Integrated Pest Management (IPM)**
- **Sampling Event data (protocols, abundance)**

NYBG

Spreadsheet vs. Database

Both used for storing and organizing data, but they differ in structure, functionality, and intended use.

	Spreadsheet	Database
Structure	Flat, two-dimensional grid	Tables with rows and columns, structured schema
Data Types	Less strict, mixed data types allowed	Predefined data types for each field
Data Relationships	Limited or manual linking	Relational data structures, tables can be linked
Data Integrity	Less robust, can be easily corrupted	Strong data integrity and consistency
Scalability	Limited, struggles with large datasets	Highly scalable, designed for large datasets
User Access	Typically single-user or limited sharing	Supports multiple users, real-time collaboration
Complexity	Easier to learn and use for basic tasks	More complex, requires some technical knowledge
Data Manipulation	Primarily manual or formula-based	Powerful querying and manipulation tools
Example	Microsoft Excel, Google Sheets	SQL Server, PostgreSQL, MySQL

Collections Management Systems (CMS)

NYBG

Collections Management System (CMS)

Database software to document and manage physical and digital objects.

Provide access to information about an instutition's collections and objects to academic researchers, staff, public.

May include ability to publish the data.

Should include built-in backup (stored in multiple locations) and recovery processes to protect data from equipment failure, disasters, human error.

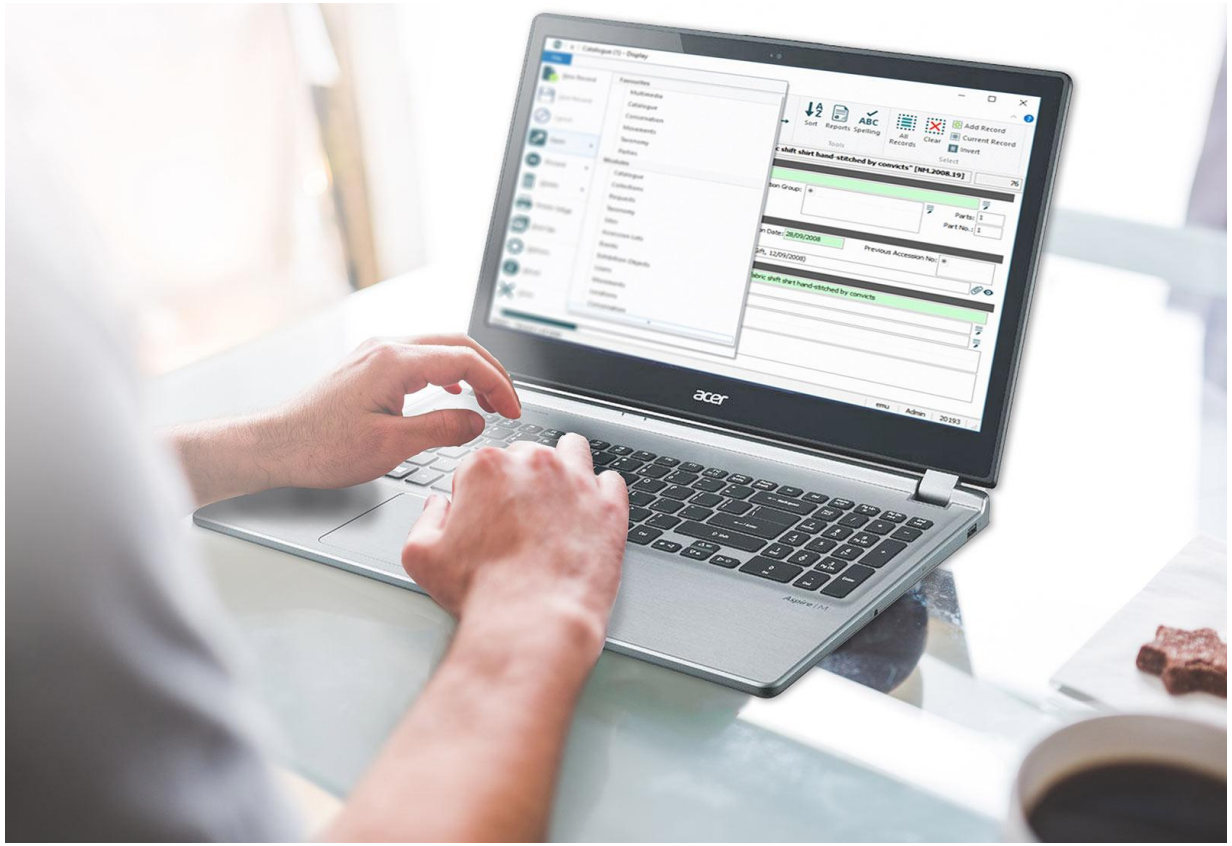


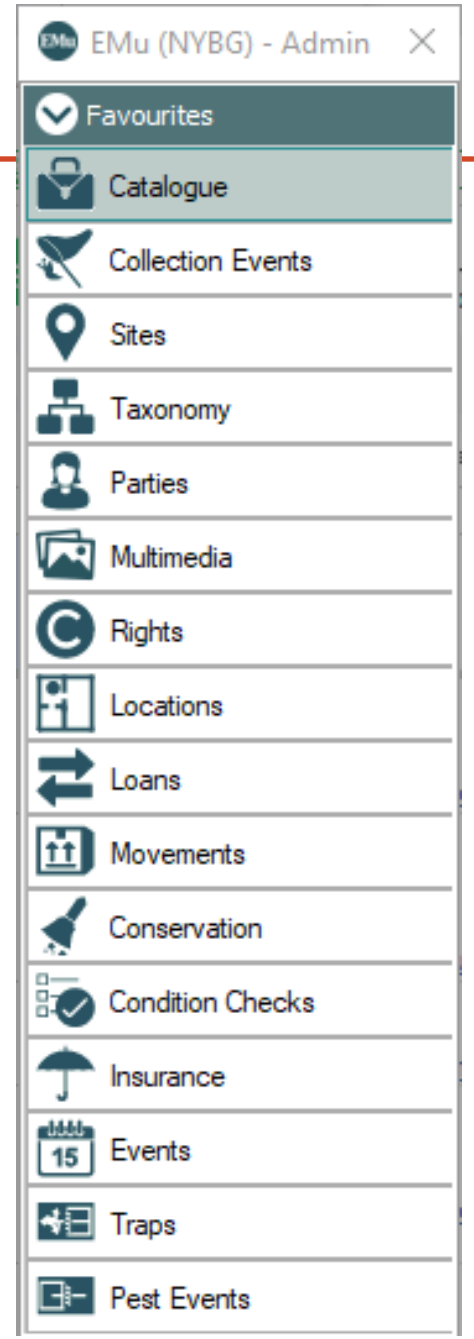
Image source: https://www.axiell.com/solutions/product/emu/

NYBG

Collections Management System (CMS)

Common features:

- **Object Catalog:** detailed information describing the object
- **Acquisition and Deaccessioning**
- **Location and Movement Control**
- **Loan management and documentation**
- **Conservation Management:** condition reports, treatment history
- **Insurance management and valuation**
- **Exhibition management**
- **Security, Rights, Permissions, Provenance**
- **Integration with Digital Asset Management System (DAMS)**



Collections Management System (CMS)

Digital Asset Management System (DAMS): database for storage, management, protection of digital assets

	CMS	DAMS
What does the software do?	Management of collections data	Management of digital assets
What does it store?	Collections data—relevant to objects, collections, exhibitions, loans, shipping, and insurance	Digital assets and related data
Which departments use it?	Collections Care Database	Collections Care Database Digital Assets Education Events Marketing Photography Publications Website
Is it difficult to use?	Not necessarily, but most museums prefer to limit access to their CMS to a group of trained key users	No, DAMS are valuable tools in working environments where many staff members require access to digital assets

Do Museums Need a DAMS? https://www.gallerysystems.com/museums-need-dams/

Collections Management System (CMS)

2013 iDigBio summary and comparison of some available products

	Database	Portal	On/Off Prem	How to Join	Cost
Arctos	Y	Y	Off	Prospective Collection Form	Annual subscription fee, Per record fee
BRAHMS	Y	Y	On	Visit OUI Software Store	Annual license fees BRAHMS Online extra
EMu	Y	Y	On or Off	Contact Axiell	Initial deployment costs Annual license & support fees
Specify	Y	Y	On or Off	Contact Specify Collections Consortium Specify code repository	Annual Membership fees
Symbiota	Y	Y	On or Off	Request to join existing portal Symbiota code repository	Collaborator on proposal Hosting fee per GB/year Portal setup fee

Arctos

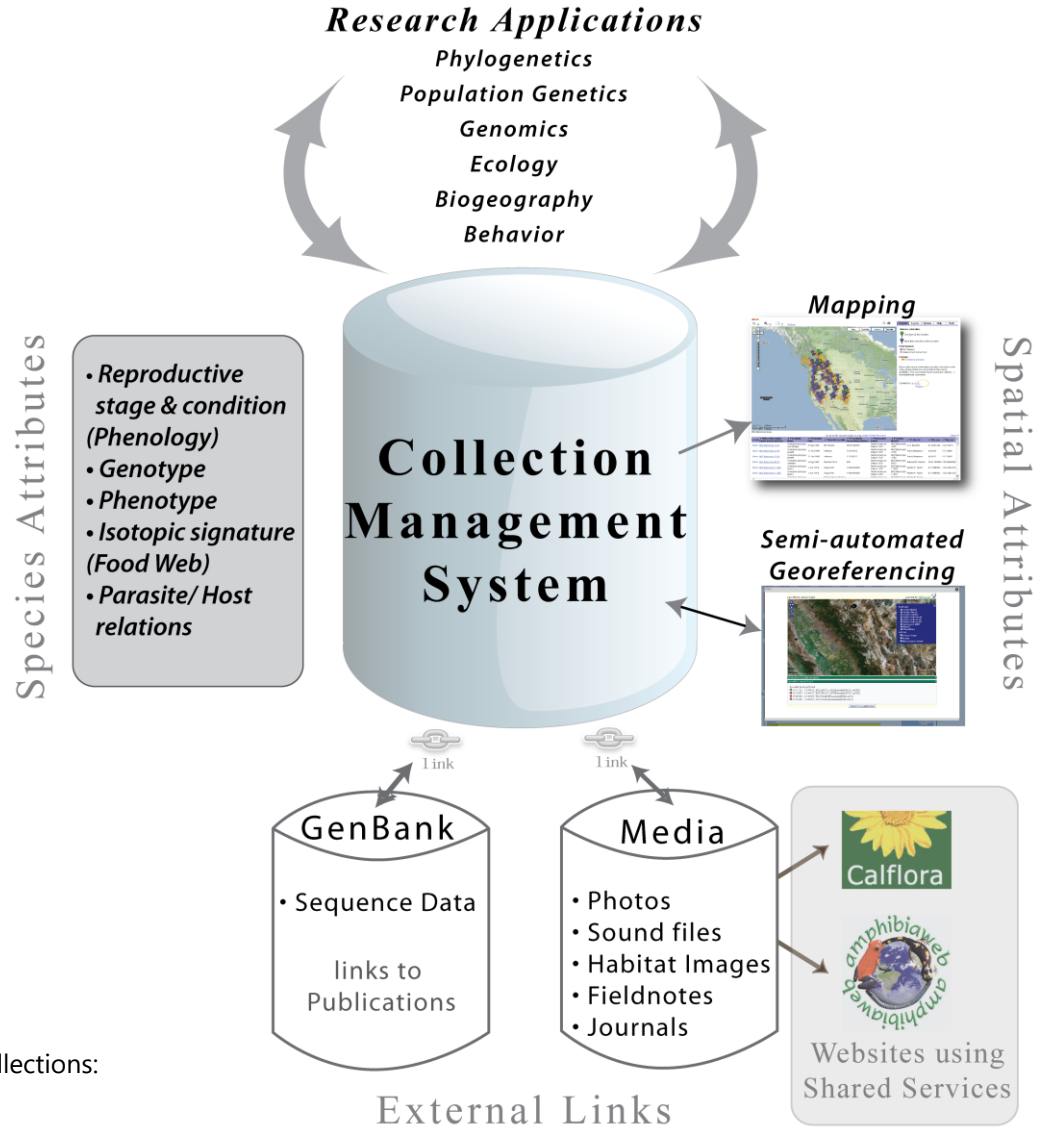
"A community and an online collection management information system… also a provider of research-grade data."

Consortium of organizations serving over 5 million records from natural and cultural history collections.

Web-based platform supporting data quality, connectivity, community

Servers hosted at Texas Advanced Computing Center, with regular backups = **no on-prem technical support needed**

Real-time edits & releases



https://arctosdb.org/about/

C. Cicero et al. (2024) Arctos: Community-driven innovations for managing natural and cultural history collections: https://doi.org/10.1371/journal.pone.0296478

NYBG

Arctos

Object-based data model with standard collections management features

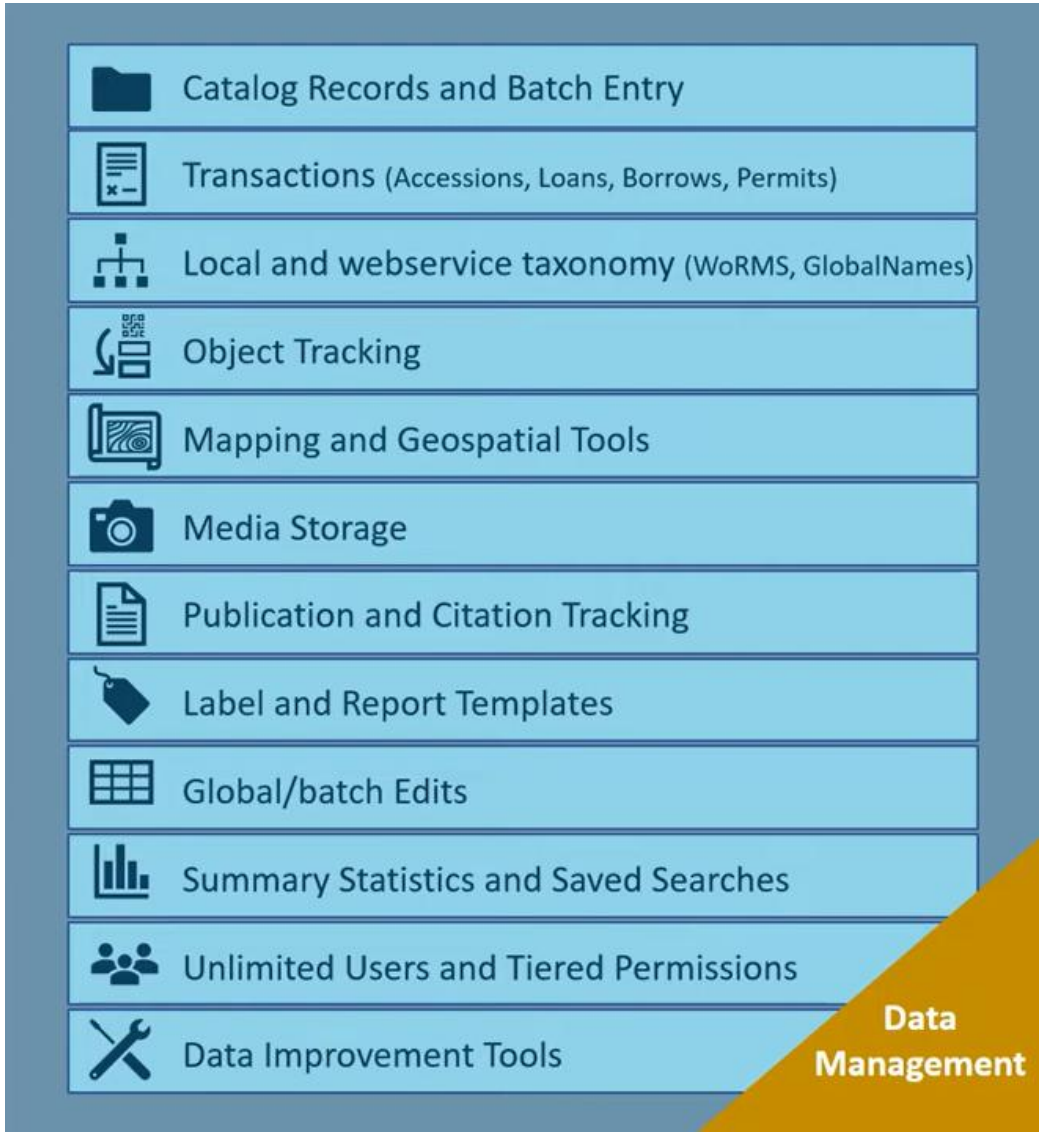
Scalable, equal access to features regardless of size

Integrated tools: geospatial, agents, annotations, conservation status, system analytics

Automated IPT

Highly active community, working groups, mentorship, user support

Video: Arctos in 10 Minutes!
https://youtu.be/6zAfzDIMV1I



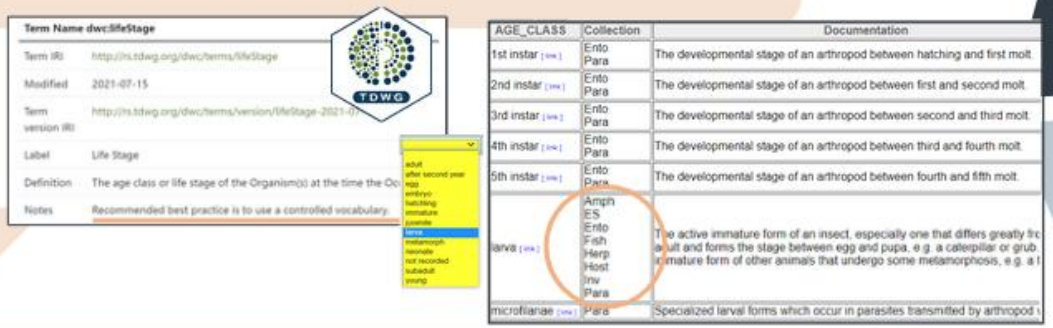
NYBG

Arctos

Shared Data Environment

Controlled vocabularies and shared authorities

Highly structured and normalized data



- **Connectivity**
- **Collaborative decision making and data improvements**
- **Automation: IPT packaging, bots, low quality/error/ ‘missed connection’ detection**

Arctos Virtual Private Database Model
with Shared Tables and Vocabularies

E. Braker (2023). Arctos Collaborative Management and Shared Data Economies.
https://docs.google.com/presentation/d/1wU8zP51bTaZaffe820JiuHeBRcPSZSdU/edit?slide=id.p3#slide=id.p3

Arctos

Annual contribution includes a subscription fee of $110, a per record fee that is tiered by the number of object records, and an administration fee. These will increase by 10% until 2026 as part of a modest adjustment over the next three years. Fees are paid to the Arctos fiscal sponsor Community Initiatives.

The per record fee is applied to collections with more than 10,000 records effective August 15, 2024 through August 15, 2025:

# RECORDS – FROM	# RECORDS – TO	PER RECORD FEE
10,001	100,000	$0.033
100,001	500,000	$0.0275
500,001	1,000,000	$0.022
1,000,001	1,500,000	$0.0165
1,500,001		$0.011

https://arctosdb.org/join-arctos/financial-contributions/

BRAHMS

Database software for managing natural history collections, botanic gardens, seed banks, field surveys, taxonomic research and biogeographic study.

https://herbaria.plants.ox.ac.uk/bol/; https://herbaria.plants.ox.ac.uk/bol/brahms/software/v8

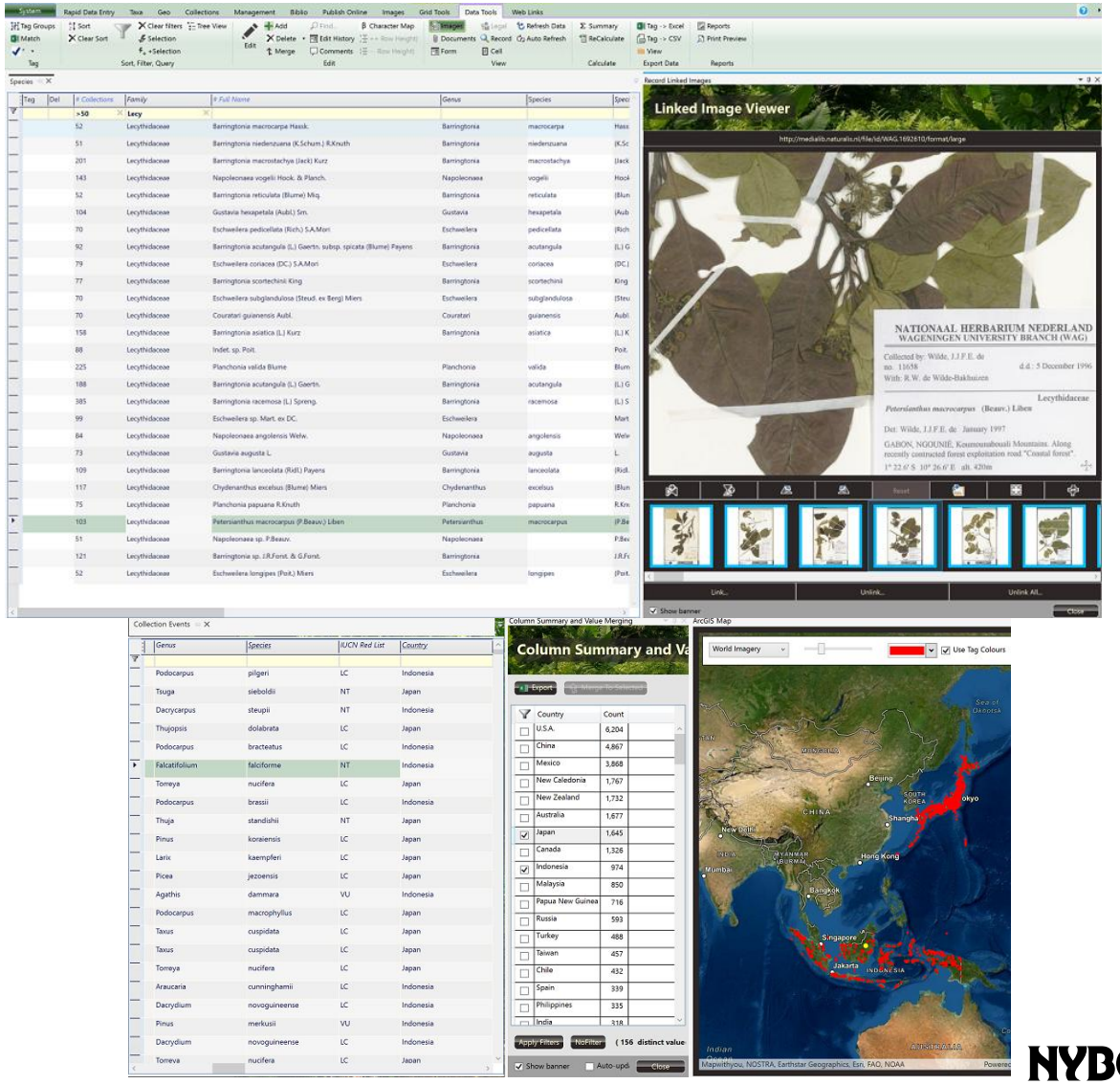
NYBG

BRAHMS

BRAHMS
Management of natural history

Key features

- Manage collections of any scale
- Networked access: multiple users edit simultaneously.
- Supports any category of preserved specimen and one to many physical specimens per collection event.
- Customizable.
- Accommodate one to many determinations, accession numbers and/or barcodes, type status, and specimen level notes.
- Link images to specimens (physical image files or media library URLs).



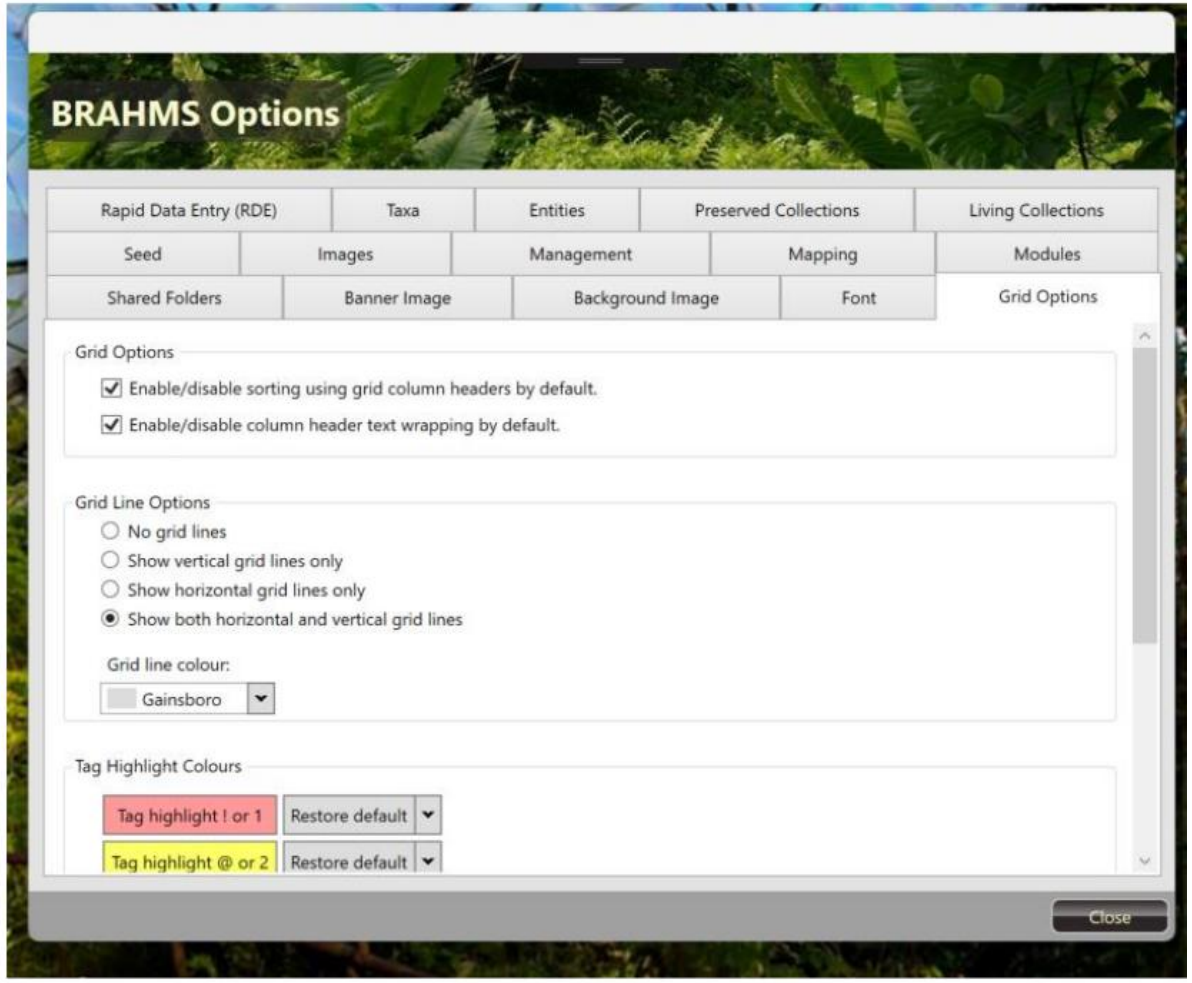
https://herbaria.plants.ox.ac.uk/bol/brahms/modules/museums;
https://herbaria.plants.ox.ac.uk/bol/brahms/software/v8

NYBG

BRAHMS

Key features (cont.)

- Design reporting templates for lists, loan forms, labels and determination slips.
- Rapid Data Entry module.
- Map geo-referenced collections. Use the map editor to locate and update map references.
- Import specimen data from other databases and websites.
- Use specimen data to develop checklists and analyze diversity for differently scaled areas.
- Publish specimen details with images online.



https://herbaria.plants.ox.ac.uk/bol/brahms/modules/museums;
https://herbaria.plants.ox.ac.uk/bol/content/software/v8/Getting_Started_BRAHMS.pdf

BRAHMS

All users require a license. License costs are broadly related to number of users.

Can purchase for permanent use with a single payment or subscribe to annual payments.

User/Organisation Type	Maximum number of concurrent users	Annual Fees
Single-site Institution	1-3	£1,500
	Add an additional user	£250 per user
Single-site Institution	Unlimited	£5,000
	Add an additional site	£1,000 per site
Multi-site Institution (single organisation)	Unlimited	£7,000

We offer a discount of 10% for licenses paid 3 years up-front.

Price list for BRAHMS Online

BRAHMS Online is an additional module with which you can design websites and publish your data and images online.

Own Server: £1000 annual fee (unlimited websites/portals)

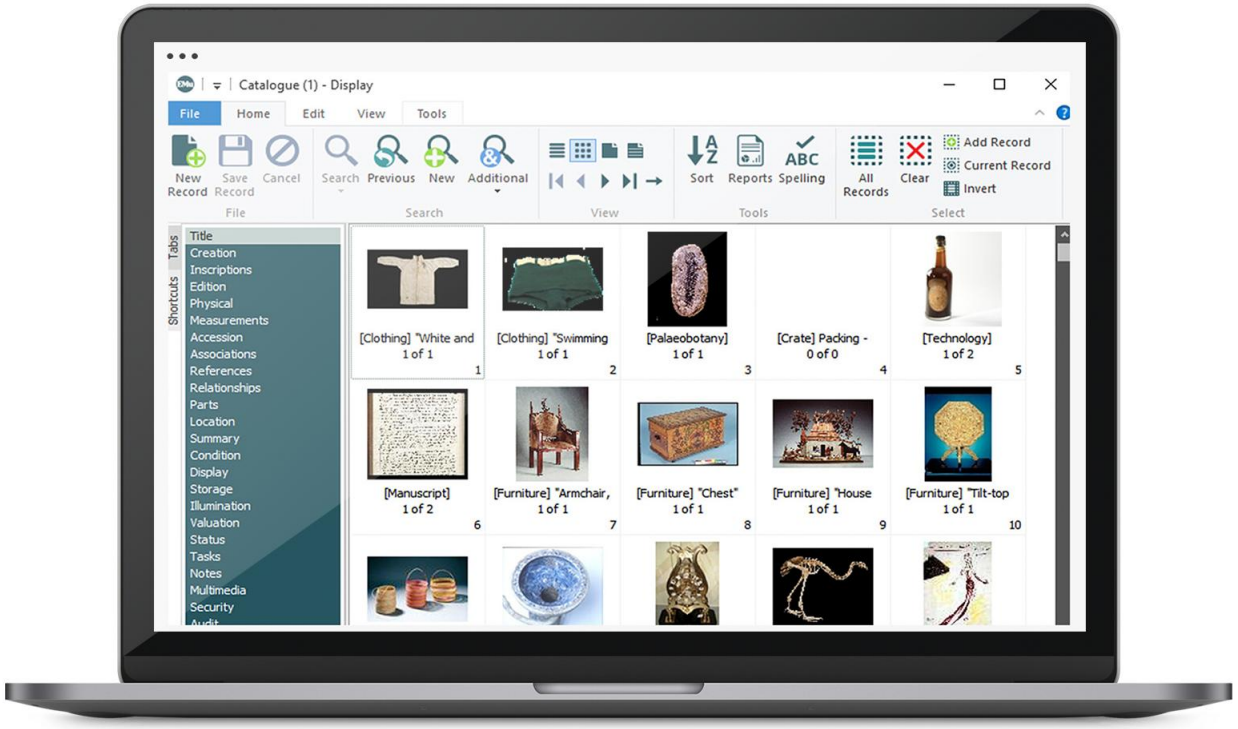
https://process.innovation.ox.ac.uk/software/p/14165/brahms/1

NYBG

EMu

Powerful, multidisciplinary CMS used by museums, art galleries, libraries, corporations, botanical gardens, herbaria

- Collections management
- Collections care, movements, IPM
- Multimedia management
- Interpretation with narratives
- Public engagement, exhibition management
- Web presence
- EMu Full Demo Video: https://www.axiell.com/solutions/emu-full-demo-video/

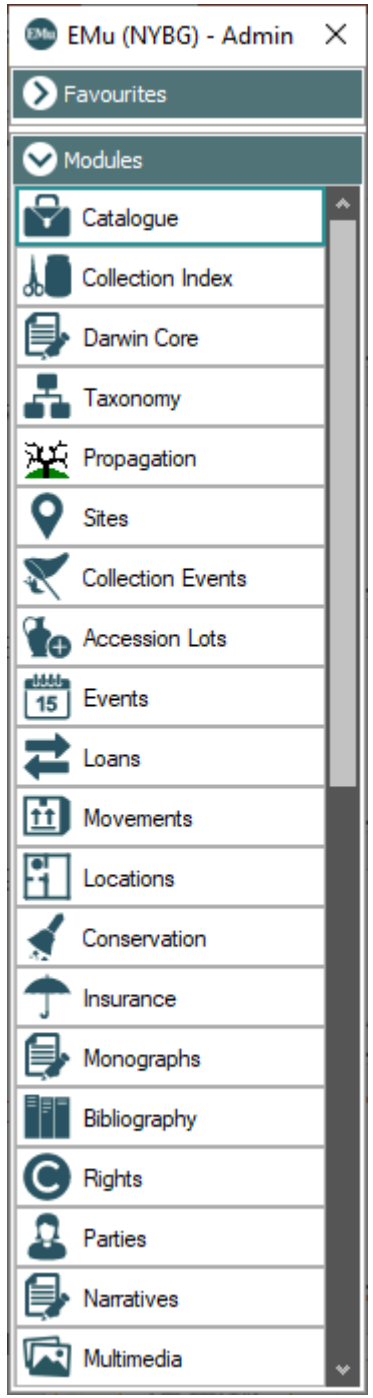


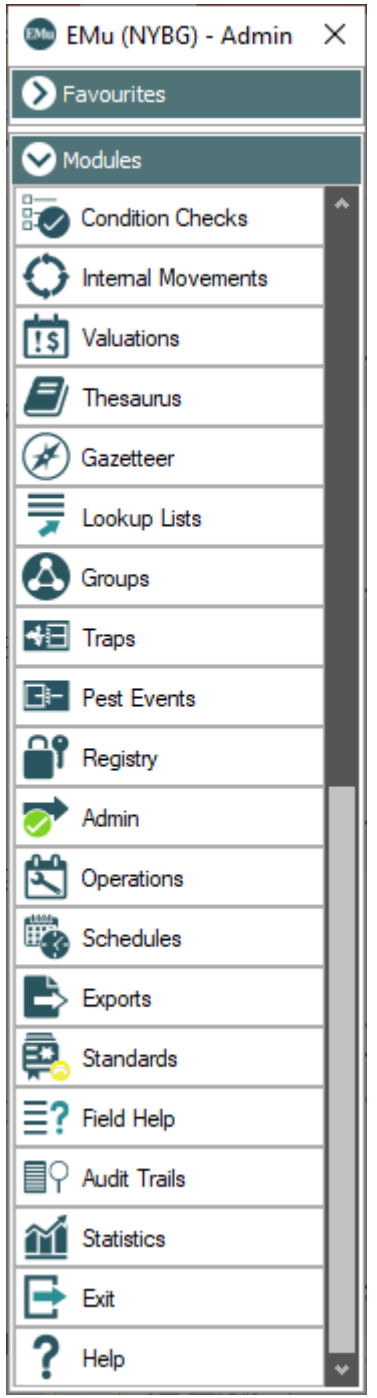
https://www.axiell.com/solutions/product/emu/
https://www.idigbio.org/wiki/images/4/4a/IDigBio-CMS-worksheet-2021-Axiell-Emu.pdf

EMu

Key Features

- Manage collections of many types, any scale.
- Networked = multiple users add/edit data simultaneously.
- Desktop client with web tools and apps (IMu, Go, Sapphire)
- Customizable, but designed so rarely needed
- Multimedia metadata management: images stored on the filesystem, EMu links to file location
- Rich XML-based import/export capabilities = interchange information with other systems
- Implement on-premise or pay Axiell to host EMu in a data center in VA.
- Integrate with DAMS, CultureConnect

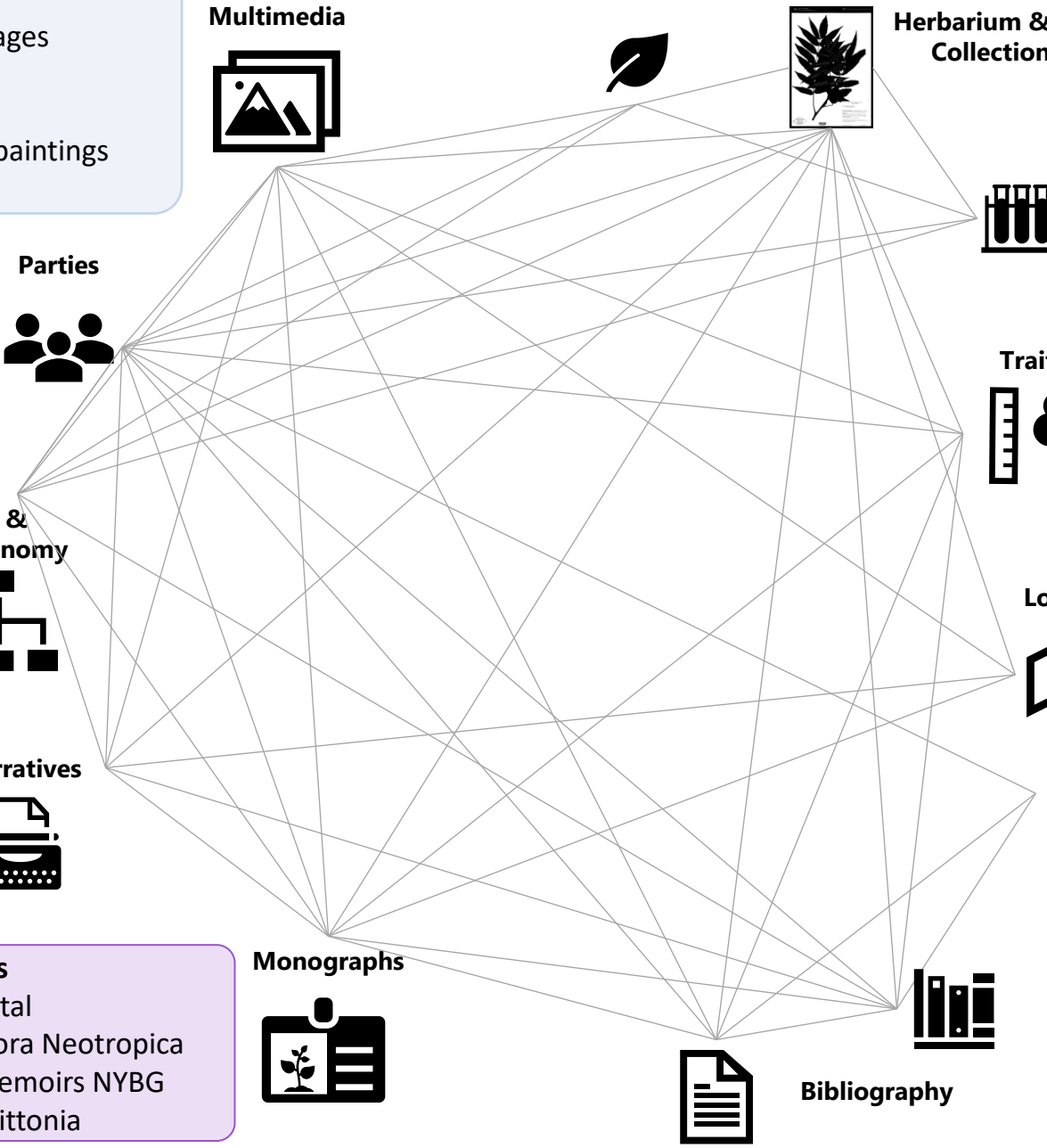




https://www.axiell.com/solutions/product/emu/

NYBG

NYBG EMu Ecosystem



Multimedia
- 3.8m Specimen images
- 100K Field images
- 6K Micrographs
- 6K Illustrations, paintings
- 15K Scanned text

Parties
- 240K Person
- 4400 Organizations
- 17K IH Person
- 4600 IH Herbarium

Taxonomy
- 1.2m Taxon names
- Plants, algae, fungi
- Internally “Filed-as”
- Externally “Accepted”
- Taxonomic hierarchy

Narratives
- 6K Total
- 600 The Hand Lens
- 3K Legume Catalog

Monographs
- 55K Total
- 10K Flora Neotropica
- 13K Memoirs NYBG
- 10K Brittonia

Catalogue
- 4.6m Herbarium specimens
- 23K Bulk Samples
- 550 DNA Aliquots
- 8K Structural Botany
- 24K Tissue samples
- 13K Use records
- 1284 with GenBank ID

Specimen Measurements or Facts
- Define, measure traits
- Link to Thesaurus Module
- Link to standardized ontologies

Sites
- 2.2m Total
- 1.2m with coordinates

Environmental Meas./Facts
- Define, measure environmental conditions
- Link to Thesaurus, ontology

Bibliography
- 5600 Books\Book Series
- 1300 Chapters
- 17K Journals
- 412K Articles

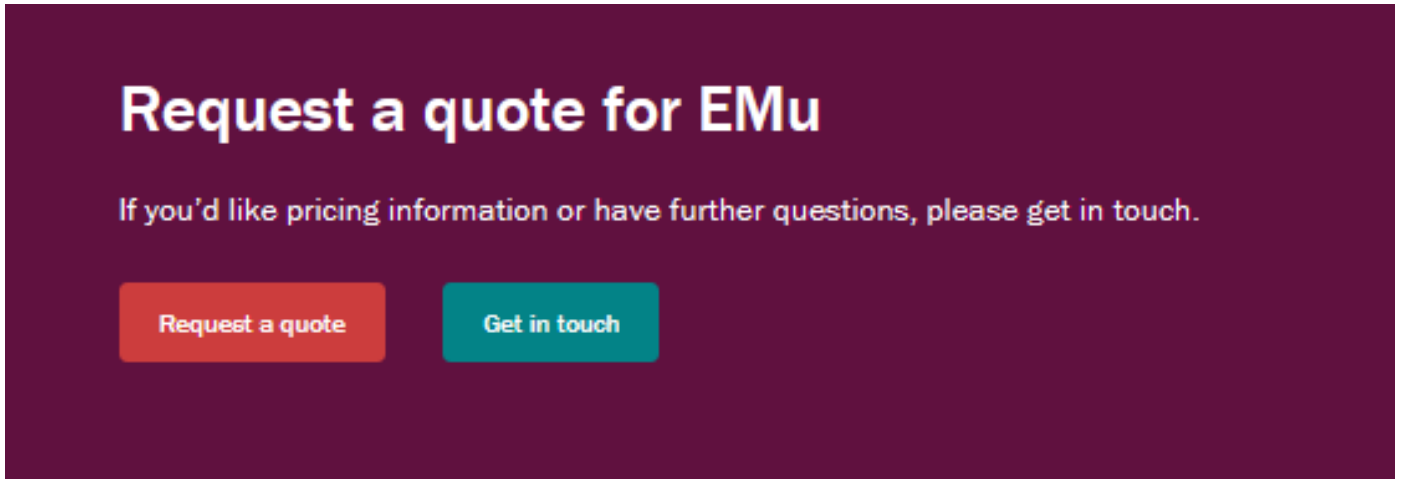
NYBG

EMu

Licensed per concurrent user. Includes all software, support, maintenance, web interface

Cost: requires a quote from Axiell.

Initial costs depends on number of licenses, extent of legacy data migration, training, customization, if Axiell will host/store data. Can be tens of thousands to hundreds of thousands of dollars.



https://collectionstrust.org.uk/software/emu/
https://www.idigbio.org/wiki/images/4/4a/IDigBio-CMS-worksheet-2021-Axiell-Emu.pdf
Detailed product brochure: https://wwwaxiellcom.cdn.triggerfish.cloud/uploads/2021/07/emubrochure2021.pdf

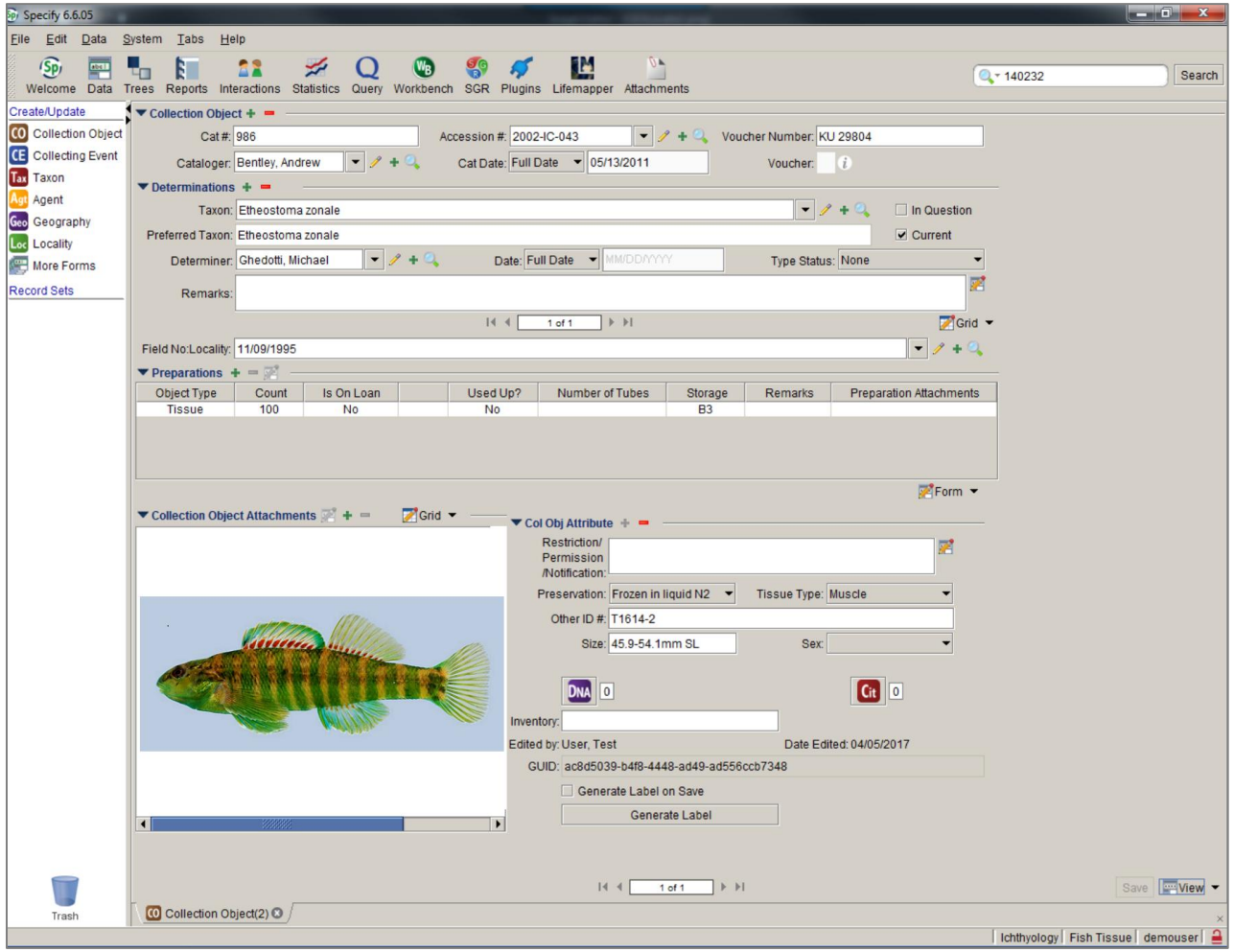
Specify

Consortium of 106 members, 300 biological collections

Collaborative, open-source software community with a vision to advance research and education uses of biological specimen and sample information

Manages specimen data, taxonomic and stratigraphic classifications, field notebooks, DNA sequence runs, literature references

Manages data associated with repository agreements, accessions, conservation treatments, collection object containers, images, document attachments



https://www.specifysoftware.org/members/

NYBG

Specify

Specify 6: Robust, full-featured desktop application

- MySQL, programmed in Java (free, open-source)

Specify 7: Specify for the web

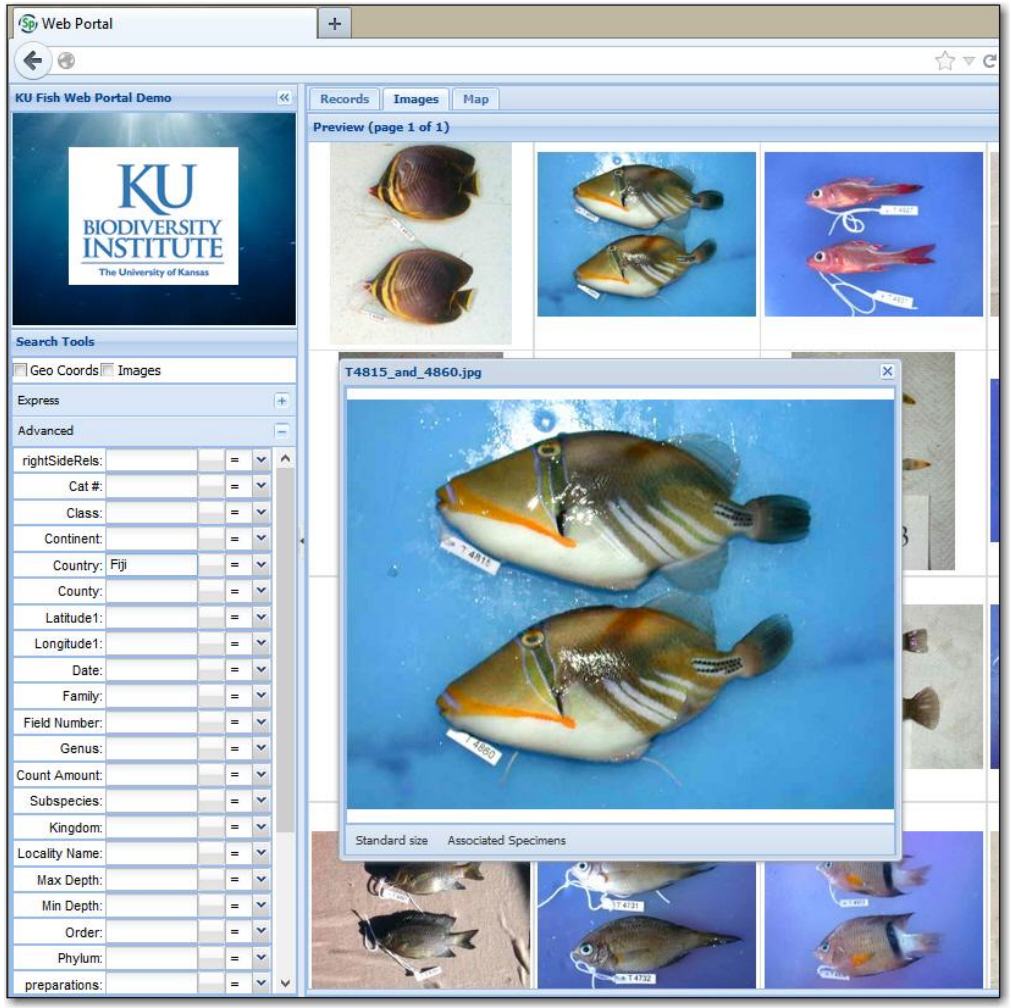
- Javascript, Linux server code in Python, same MySQL
- Run simultaneously with Specify 6 (same MySQL)

Specify Cloud Hosting

- They install and host Specify 7 in the cloud

Add Ons:

- **Attachment Server:** image storage; for web portal
- **Web Portal:** read-only, public access to data and images



https://www.specifysoftware.org/products/; https://www.specifysoftware.org/add-ons/

R. J. Burkhalter. Specify as a database management tool. https://www.idigbio.org/wiki/images/4/4a/Roger_Specify_talk.pdf

NYBG

Specify

Specify COLLECTIONS CONSORTIUM

Welcome to the Speciforum!

Members of the **Specify Collections Consortium** can join the conversation! Please read **our guidelines** before you begin!

create a topic

ask a question

docs

email support

Get Help

This is the place for raising and discussing any specific issue, including workflows, tasks, data wrangling, software issues, and getting things done.

Configuration & Installation 191 Topics	**WorkBench Workflows** 43 Topics	**Data Entry** 41 Topics	**Documentation** 30 Topics
Query Building 23 Topics	**Labels & Reports** 23 Topics	**Trees** 17 Topics	**Forms** 12 Topics
Interactions 10 Topics	**Specify 6** 274 Topics	**Specify 7** 204 Topics	**Videos** 108 Topics
Webinar 31 Topics	**Workshop** 11 Topics		

https://discourse.specifysoftware.org/

NYBG

Specify

Founding Partner: $40,000/year/Institution (all institution collections)

- 2 permanent seats on Board, 1 seat on each of 2 committees; **direct access to Consortium engineers for assistance; priority access to technical support services;** collaborative proposal/project planning for software development, data analysis, integration; **priority access to new products/updates**; 2 paid seats at annual Summit

Full Member: $5000/year/collection

- Access to rotating seat on the Board, priority for seat on committee**; priority access to technical support services**; collaborative proposal/project planning; **preferred access to new products/updates**; 1 paid seat at annual Summit

Solutions Member: $1250-3000/year/collection

- Access to rotating seats on committees**; access to technical support services**; technical advice for project planning; **timely access to all tools and updates**; 1 paid seat at annual Summit

Associate Member: $1000/year/collection

- Access to rotating seats on committees; **full access to Specify forum**; technical advice for funding proposal/project; **access to all tools and updates**; 1 paid seat at annual Summit

https://www.specifysoftware.org/membership-levels/#levels

NYBG

Symbiota

Open-source software for managing and mobilizing biodiversity data.

Specifically designed toward efficient, collaborative digitization and an open data exploration and publishing tool.

Distributed network of theme-based research portals incorporating data from >1900 collections

60+ portals publish over 95m occurrences and 43m images

Researchers, educators, explorers, and more

Anyone can explore and use biodiversity data stored in Symbiota-based portals. Symbiota portals include tools for:

- species inventories
- interactive identification keys
- integrated specimen and field images
- taxonomic information
- species distribution maps
- taxonomic descriptions
- public datasets

https://symbiota.org/

NYBG

Symbiota

1. **Deploy and maintain your own portal** using open-source code and your own IT infrastructure
    - Linux or Windows, PHP enabled web server, MariaDB
    - Example: Oregon State University Herbarium
2. **Request to join an existing portal**

Dataset Management

- **Live Manage:** catalog and manage collections directly within portal interface
- **Snapshot:** upload/update a static dataset from another database, but do not manage data directly within the portal interface

Collectors, collection managers, and curators

Content can be actively managed in Symbiota portals. Features include:

- data entry from label images
- data harvesting from specimen duplicates
- batch georeferencing (even across collections)
- data validation and cleaning
- generating progress reports
- loan management
- data publishing to GBIF and iDigBio

https://symbiota.org/

NYBG

Symbiota

Existing Portals:

- Taxonomic and/or geographic in Scope
- May be hosted by Symbiota Support Hub (SSH provides technical infrastructure)
- May be part of the SEINet Portal Network (11 portals, 1 shared database)
- May be maintained by specific institution

Portals Campaigns: month-long SSH support for specific portal to help with updating metadata, publishing, data quality, training, expanding community.

Resources: Help Desk, Documentation, Video Tutorials, GitHub Discussions, Example Workflows

NYBG

Symbiota

Sustaining Symbiota Services:

1. **iDigBio Symbiota Support Hub:** funded to 08/2026 (5-year iDigBio award) to support portals that originated as TCNs and PENs prior to 2021. Does not provide for new development or for portals created outside of TCNs and PENs or after 2020.

2. **Other Symbiota Research Awards:** for all services outside #1, Symbiota Team is able to collaborate on proposals for NSF, USDA, NIH. **Strongly prefer to collaborate to act as equal academic partners**, which can support new development, data management, IT infrastructure, and other scientific collaborations.

3. **Fee-Based Symbiota Service Contracts:** since 2023, KU Symbiota (KU Service Center) provides quotes and invoices to cover full spectrum of Symbiota services. Act as service providers, not equal academic partners. Operate under a cost recovery model.

    - **Hosting: Data Portals & Images** = $2/GB/year (nonprofit discounts available)
    - **Data Portal Setup** = $3583/portal (full setup)

https://symbiota.org/sustaining-symbiota-services/

NYBG

Collections Management System (CMS)

2013 iDigBio summary and comparison of some available products

	Database	Portal	On/Off Prem	How to Join	Cost
Arctos	Y	Y	Off	Prospective Collection Form	Annual subscription fee, Per record fee
BRAHMS	Y	Y	On	Visit OUI Software Store	Annual license fees BRAHMS Online extra
EMu	Y	Y	On or Off	Contact Axiell	Initial deployment costs Annual license & support fees
Specify	Y	Y	On or Off	Contact Specify Collections Consortium Specify code repository	Annual Membership fees
Symbiota	Y	Y	On or Off	Request to join existing portal Symbiota code repository	Collaborator on proposal Hosting fee per GB/year Portal setup fee

Collections Management System (CMS)

When choosing a CMS, perform a needs assessment:

- Identify users
- Know your data
- Know your collections
- Determine where and how a CMS can increase efficiency
- Decide where data will be stored
- Know your budget
- Seek support from others who have been through the process

HAI The Research, Discovery, and Experience Company: Start your Museum CMS Selection Process with a Needs Assessment

Collections Management System (CMS)

In preparing for the new CMS, evaluate:

- **Data Quality:** data for each object is accurate, complete, located in appropriate fields, follows best practices
- **Data Format:** data schema, data fields, data export
- **Data Location:** spreadsheet, unique database, CMS on-premise, CMS on the cloud
- **Digital Asset Format & Location:** local computer or server, cloud service, in a separate system

Woody, R. C. (2023). How to Prepare for a New Museum CMS. LUCIDEA. https://lucidea.com/blog/how-to-prepare-for-a-new-museum-cms/
Woody, R. C. (2019). How to Prepare for a Museum Collections Management System Migration. LUCIDEA. https://lucidea.com/blog/how-to-prepare-for-a-museum-collections-management-system-migration/

Biodiversity Data Standards

NYBG

Data Standards

"The whole is greater than the sum of its parts" - Aristotle

Aggregating biodiversity data managed across many different databases (and schemas) requires standards and guidelines for the recording and exchanging of those data.

Taxonomic Databases Working Group, now Biodiversity Information Standards (TDWG)

- Non-profit organization and a community
- International collaboration among creators, managers, and users of biodiversity information to promote wider more effective dissemination
- Develop, ratify, promote standards and guidelines
- Act as a forum for discussing all aspects of biodiversity information management

https://www.tdwg.org/

Data Standards

Many current and actively maintained TDWG standards

Darwin Core:

- Terms to facilitate sharing information about biodiversity

Audiovisual Core:

- Terms to represent metadata for biodiversity multimedia resources and collections

https://www.tdwg.org/standards/; https://www.tdwg.org/standards/dwc/; https://www.tdwg.org/standards/ac/

Darwin Core List of Terms

Biodiversity Information Standards TDWG

3 Term indices

3.1 Index By Term Name

(See also 3.2 Index By Label)

Classes

dwc:Dataset | dwc:Event | dwc:EventAttribute | dwc:EventMeasurement | dwc:FossilSpecimen | dwc:GeologicalContext | dwc:HumanObservation | dwc:Identification | dwc:LivingSpecimen | dcterms:Location | dwc:MachineObservation | dwc:MaterialCitation | dwc:MaterialEntity | dwc:MaterialSample | dwc:MeasurementOrFact | dwc:Occurrence | dwc:OccurrenceMeasurement | dwc:Organism | dwc:PreservedSpecimen | dwc:ResourceRelationship | dwc:Sample | dwc:SampleAttribute | dwc:SamplingEvent | dwc:SamplingLocation | dwc:Taxon

Record level

dwc:accordingTo | dwc:accuracy | dwc:basisOfRecord | dwc:collectionCode | dwc:collectionID | dwc:dataGeneralizations | dwc:datasetID | dwc:datasetName | dwc:DwCType | dwc:dynamicProperties | dwc:Generalizations | dwc:informationWithheld | dwc:institutionCode | dwc:institutionID | dwc:ownerInstitutionCode

On this page

Darwin Core Quick Reference Guide

Biodiversity Information Standards TDWG

Easy-to-read reference with term names, definitions, examples and comments

Occurrence

occurrenceID | catalogNumber | recordNumber | recordedBy | recordedByID | individualCount | organismQuantity
organismQuantityType | sex | lifeStage | reproductiveCondition | caste | behavior | vitality | establishmentMeans
degreeOfEstablishment | pathway | georeferenceVerificationStatus | occurrenceStatus | associatedMedia
associatedOccurrences | associatedReferences | associatedTaxa | otherCatalogNumbers | occurrenceRemarks

recordedBy	
Identifier	http://rs.tdwg.org/dwc/terms/recordedBy
Definition	A list (concatenated and separated) of names of people, groups, or organizations responsible for recording the original dwc:Occurrence. The primary collector or observer, especially one who applies a personal identifier (dwc:recordNumber), should be listed first.
Comments	Recommended best practice is to separate the values in a list with space vertical bar space (|). This term has an equivalent in the dwciri: namespace that allows only an IRI as a value, whereas this term allows for any string literal value.
Examples	`José E. Crespo` `Oliver P. Pearson | Anita K. Pearson` (where the value in recordNumber `OPP 7101` corresponds to the collector number for the specimen in the field catalog of Oliver P. Pearson)

On this page

Darwin Core Maintenance Group. 2021. Darwin Core Quick Reference Guide. Biodiversity Information Standards (TDWG). https://dwc.tdwg.org/terms/

NYBG

Darwin Core List with Term Version

Biodiversity Information Standards TDWG

Term Name dwc:recordedBy

Term IRI	http://rs.tdwg.org/dwc/terms/recordedBy
Modified	2023-06-28
Term version IRI	http://rs.tdwg.org/dwc/terms/version/recordedBy-2023-06-28
Label	Recorded By
Definition	A list (concatenated and separated) of names of people, groups, or organizations responsible for recording the original dwc:Occurrence. The primary collector or observer, especially one who applies a personal identifier (dwc:recordNumber), should be listed first.
Notes	Recommended best practice is to separate the values in a list with space vertical bar space (|). This term has an equivalent in the dwciri: namespace that allows only an IRI as a value, whereas this term allows for any string literal value.
Examples	`José E. Crespo` `Oliver P. Pearson | Anita K. Pearson` (where the value in recordNumber `OPP 7101` corresponds to the collector number for the specimen in the field catalog of Oliver P. Pearson)

TDWG

- GitHub

Metadata for the 2023-06-28 version of the term dwc:recordedBy

Term Name:	dwc:recordedBy
Label:	Recorded By
Term version IRI:	http://rs.tdwg.org/dwc/terms/version/recordedBy-2023-06-28
Version of:	http://rs.tdwg.org/dwc/terms/recordedBy
Issued:	2023-06-28
Definition:	A list (concatenated and separated) of names of people, groups, or organizations responsible for recording the original dwc:Occurrence. The primary collector or observer, especially one who applies a personal identifier (dwc:recordNumber), should be listed first.
Type:	Property
Status:	recommended
Replaces:	http://rs.tdwg.org/dwc/terms/version/recordedBy-2017-10-06

Darwin Core Maintenance Group. 2021. Darwin Core Quick Reference Guide. Biodiversity Information Standards (TDWG). https://dwc.tdwg.org/terms/

NYBG

Darwin Core Q & A Wiki

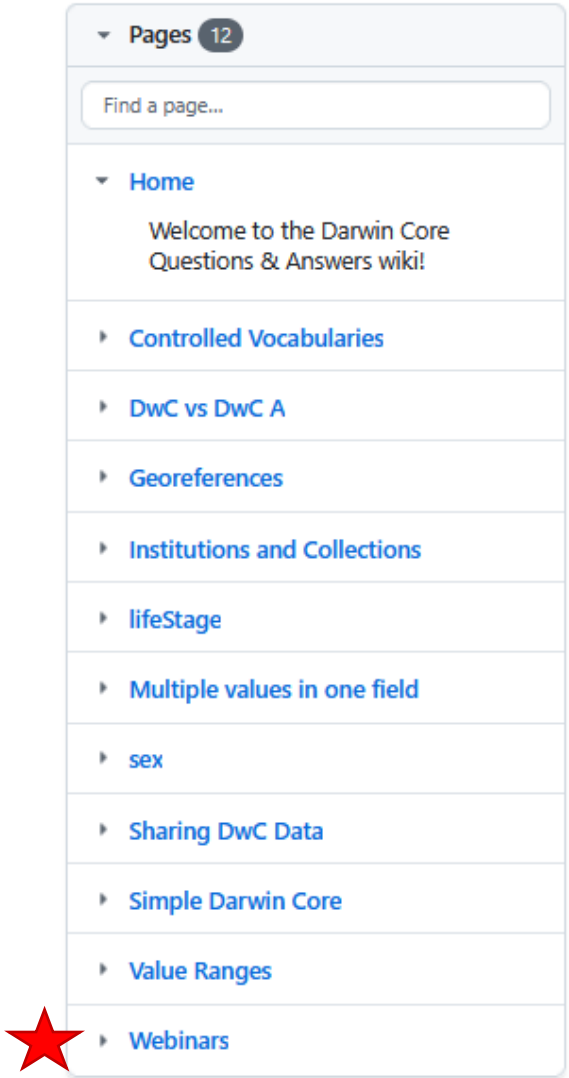
Biodiversity Information Standards TDWG

Access to the outcomes of discussions about the use of Darwin Core

List of pages on broad topics or use the Search box

Darwin Core Hour series

Announcement	Slides	Video	Adobe Connect Recording	Chapter Abstracts
Introduction to Darwin Core Hour	Slides	Vimeo	2017-02-07	Chapter 0
Introduction to Darwin Core	Slides	Vimeo	2017-02-07	Chapter 1
Even Simple is Hard	Slides	Vimeo	2017-03-07	Chapter 2
Thousands of Shades for "Controlled" Vocabularies	Slides	Vimeo	2017-04-04	Chapter 3
Evolution of Darwin Core Terms and Extensions - two extant examples for community input. Part 1. Preparations Part 2. occurrenceStatus and establishmentMeans	Part1 Part2	Vimeo	2017-05-02	Chapter 4



https://github.com/tdwg/dwc-qa/wiki

NYBG

Audiovisual Core

Biodiversity Information Standards TDWG

Audiovisual Core Multimedia Resources Metadata Schema = Audiovisual Core (AC)

TDWG data standard including a set of vocabularies designed to represent metadata for biodiversity multimedia resources and collections; a schema for exchanging these metadata

Term Name: dcterms:identifier	
Normative URI	http://purl.org/dc/terms/identifier
Modified	2020-01-27
Term version URI	http://dublincore.org/usage/terms/history/#identifierT-001
Label	Identifier
★	**Required:** Yes for media collections, No for media resources (but preferred if available) -- **Repeatable:** Yes
Definition	An unambiguous reference to the resource within a given context.
Usage	An arbitrary code that is unique for the resource, with the resource being either a provider, collection, or media item.
Notes	Using multiple identifiers implies that they have a same-as relationship, i.e. they all identify the same object (e. g. an object may have all of an http-URL, an lsid-URI, and a UUID).

7 Vocabularies

- 7.1 Management Vocabulary
- 7.2 Attribution Vocabulary
- 7.3 Agents Vocabulary
- 7.4 Content Coverage Vocabulary
- 7.5 Geography Vocabulary
- 7.6 Temporal Coverage Vocabulary
- 7.7 Taxonomic Coverage Vocabulary
- 7.8 Resource Creation Vocabulary
- 7.9 Related Resources Vocabulary
- 7.10 Service Access Point Vocabulary
- 7.11 Region of Interest Vocabulary

https://www.tdwg.org/standards/ac/
https://ac.tdwg.org/termlist/

NYBG

Authority Files and Controlled Vocabularies

NYBG

Authority Files

Database tables containing authority records representing e.g., names of people, places, things, and concepts

- Standardized, authoritative
- Provide data control, standardization, consistency, cataloguing efficiency
- Attach to other records within other tables and/or the same table

Identifiers

- Database identifier unique within the database
- Globally Unique Identifiers (GUIDs) generated for external use
- Identifiers or GUIDs incorporated from external resources

Authority Files

PARTIES

- Person
- Organization
- Collaboration
- Position
- IH Herbarium
- IH Person

TAXONOMY

- Names
- Hierarchy
- Synonymy

BIBLIOGRAPHY

- Publications
- Hierarchy
- BHL sync

THESAURUS

- Definitions
- Hierarchy
- Synonymy
- Import
- Validation

LOOKUP LISTS

- Control vocab
- Hierarchy
- ReadOnly

NYBG

NYBG EMu Ecosystem

Multimedia
- 3.8m Specimen images
- 100K Field images
- 6K Micrographs
- 6K Illustrations, paintings
- 15K Scanned text

Multimedia

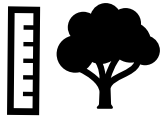
Herbarium & Lab Collections

Catalogue
- 4.6m Herbarium specimens
- 23K Bulk Samples
- 550 DNA Aliquots
- 8K Structural Botany
- 24K Tissue samples
- 13K Use records
- 1284 with GenBank ID

Parties
- 240K Person
- 4400 Organizations
- 17K IH Person
- 4600 IH Herbarium

Parties

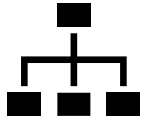
Traits



Specimen Measurements or Facts
- Define, measure traits
- Link to Thesaurus Module
- Link to standardized ontologies

Taxonomy
- 1.2m Taxon names
- Plants, algae, fungi
- Internally “Filed-as”
- Externally “Accepted”
- Taxonomic hierarchy

Taxa & Taxonomy



Localities

Sites
- 2.2m Total
- 1.2m with coordinates

Environmental Meas./Facts
- Define, measure environmental conditions
- Link to Thesaurus, ontology

Narratives
- 6K Total
- 600 The Hand Lens
- 3K Legume Catalog

Monographs
- 55K Total
- 10K Flora Neotropica
- 13K Memoirs NYBG
- 10K Brittonia

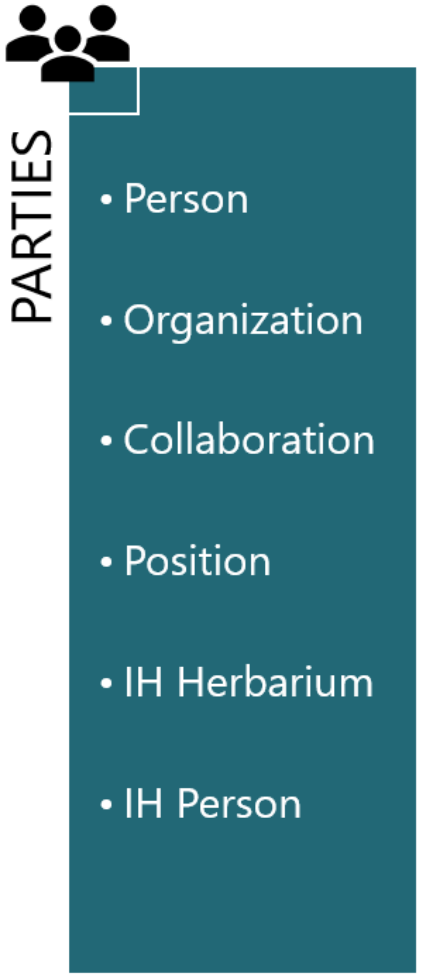
Monographs

Bibliography

Bibliography
- 5600 Books\Book Series
- 1300 Chapters
- 17K Journals
- 412K Articles

NYBG

Parties or Agents



Records represent individuals/organizations associated with collections/organization

- **Person:** collectors, authors, determiners, researchers, staff, donors, etc.
- **Organization:** research institutions, funding agencies, etc.

Information to disambiguate: address, roles, associations, biography, synonymy

- Avoid sensitive information
- Use caution when publishing data online

Import existing data from authoritative source

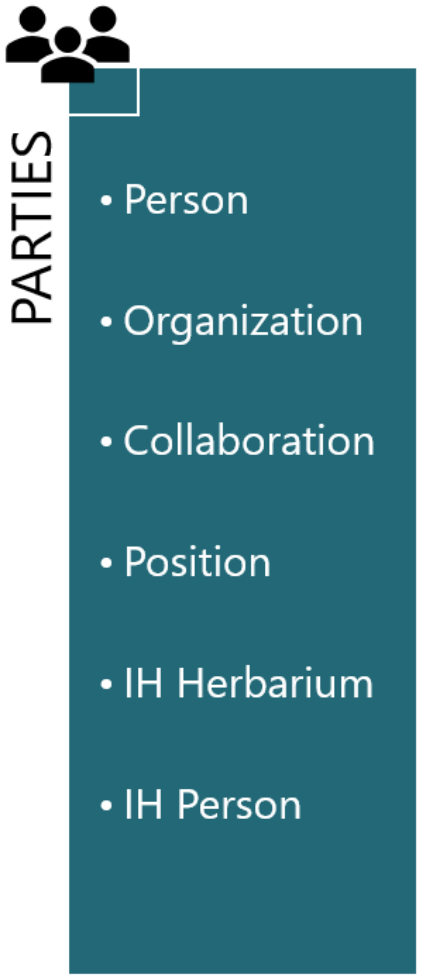
Incorporate identifiers from (link to) open resources: ORCID, Wikidata, Bionomia

Recommended reading:

- People are essential to linking biodiversity data. Groom et al. 2020, https://doi.org/10.1093/database/baaa072
- The disambiguation of people names in biological collections. Groom et al. 2022, https://doi.org/10.3897/BDJ.10.e86089

NYBG

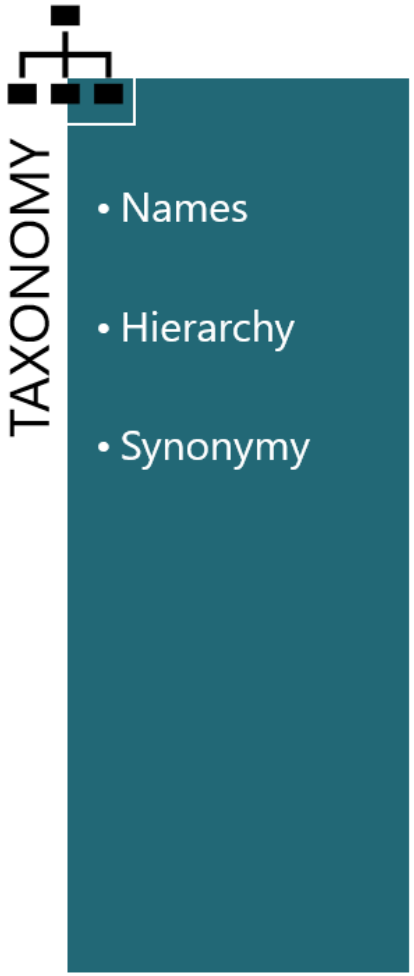
Parties or Agents



Online Resources for Parties/Agents

- ORCID
- Wikidata
- Bionomia
- HUH Index of Botanists
- IPNI Index of Authors (Advanced search or see full list here)
- Index Herbariorum

Taxonomy

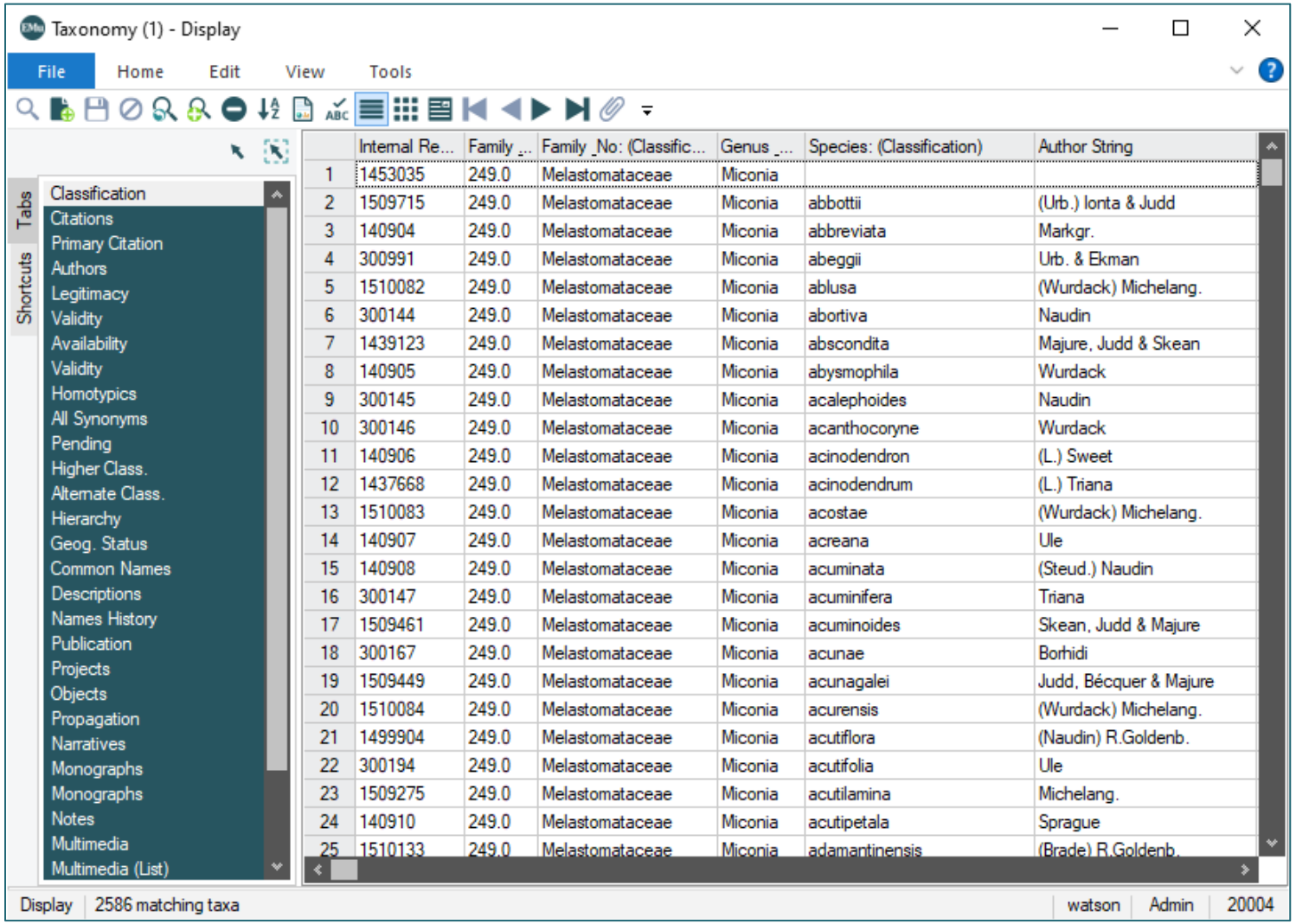


Records with detailed information about taxonomic names relating to a collection

- **Full hierarchy:** Kingdom to infraspecies
- **Can include:**
  - Relationships between names (synonymies)
  - Protologue citations, types specimens
  - CITES, IUCN conservation status (opt to publish or withhold occurrence locality)
- **Identifiers:**
  - Database-assigned identifier
  - Generate GUIDs for external use
  - Incorporate GUIDs from external resources

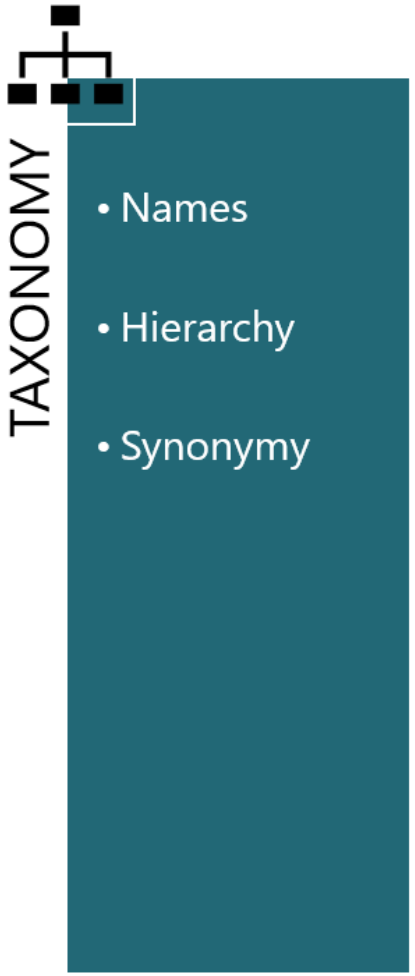
Taxonomy

- Populate with cleaned, standardized scientific names represented in the collections
- Utilize taxonomic authority databases
- Incorporate one to many taxon identifiers from authority databases
- Records may attach to other records in Taxonomy and many other tables



NYBG

Taxonomy



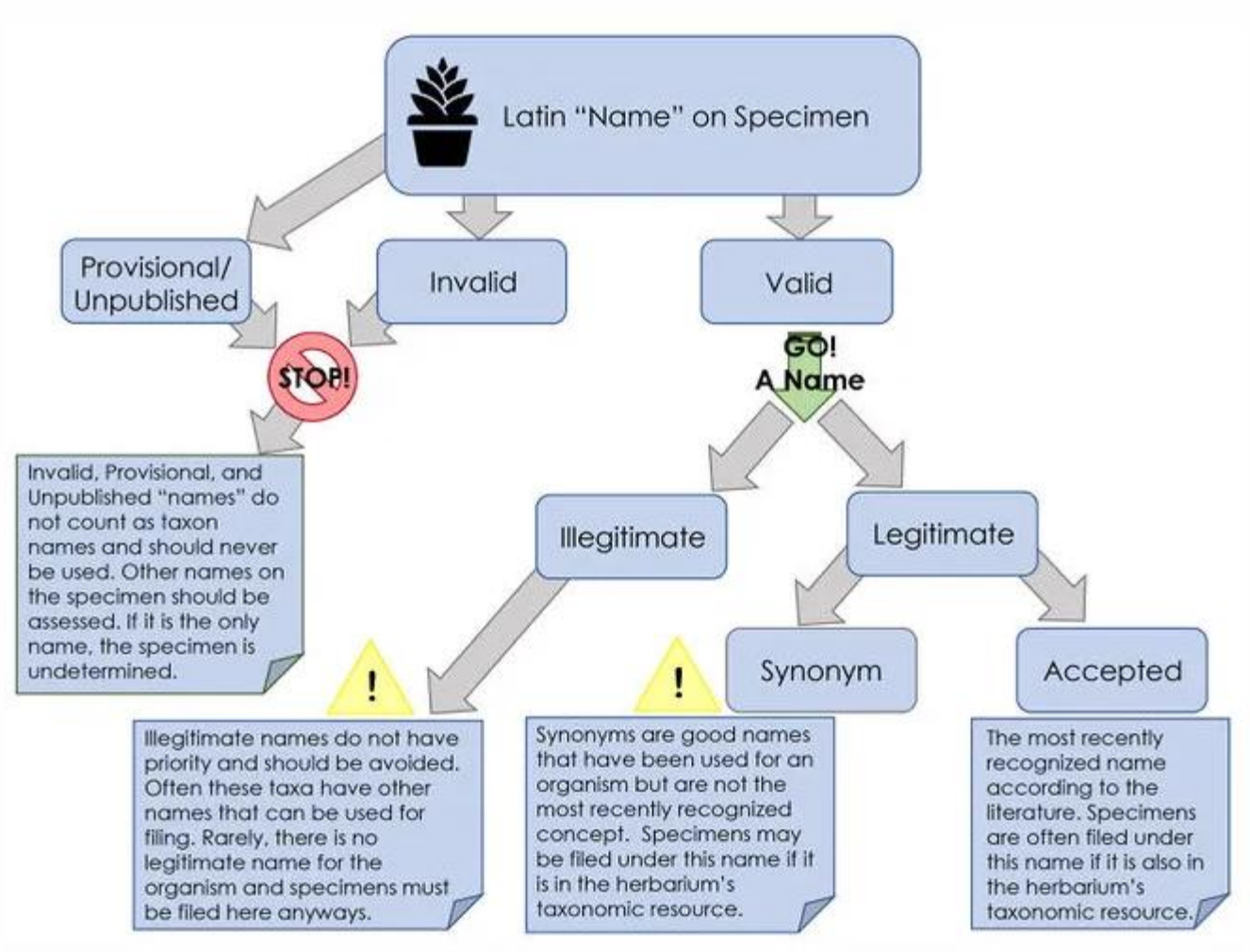
Online Resources for Taxonomy

- Plants of the World Online = International Plant Names Index (IPNI) = World Checklist of Vascular Plants (WCVP)
- World Flora Online
- Integrated Taxonomic Information System (ITIS), Compare Taxonomy/Nomenclature
- Taxonomic Names Resolution Service
- TROPICOS
- GBIF
- Angiosperm Phylogeny Website
- Index Fungorum
- MycoBank
- AlgaeBase
- Index Nominum Algarum
- World Register of Marine Species
- SPNHC wiki

Taxonomy

[SPNHC Taxonomic Resources]

- Name(s) on a specimen may not be published, legitimate, or valid.
- Each collection may have different guidelines for cataloging and handling taxonomic names
- Recommended to never use invalid, provisional, and unpublished "names." If no other names present to assess, specimen is undetermined.



Flowchart to determine the type of name on a botany specimen

Image from https://spnhc.org/taxonomic-resources/#Botanical_Resources

NYBG

Taxonomy Identifiers

SPECIES | ACCEPTED

Miconia petroniana Cogn. & Saldanha ex Cogn.

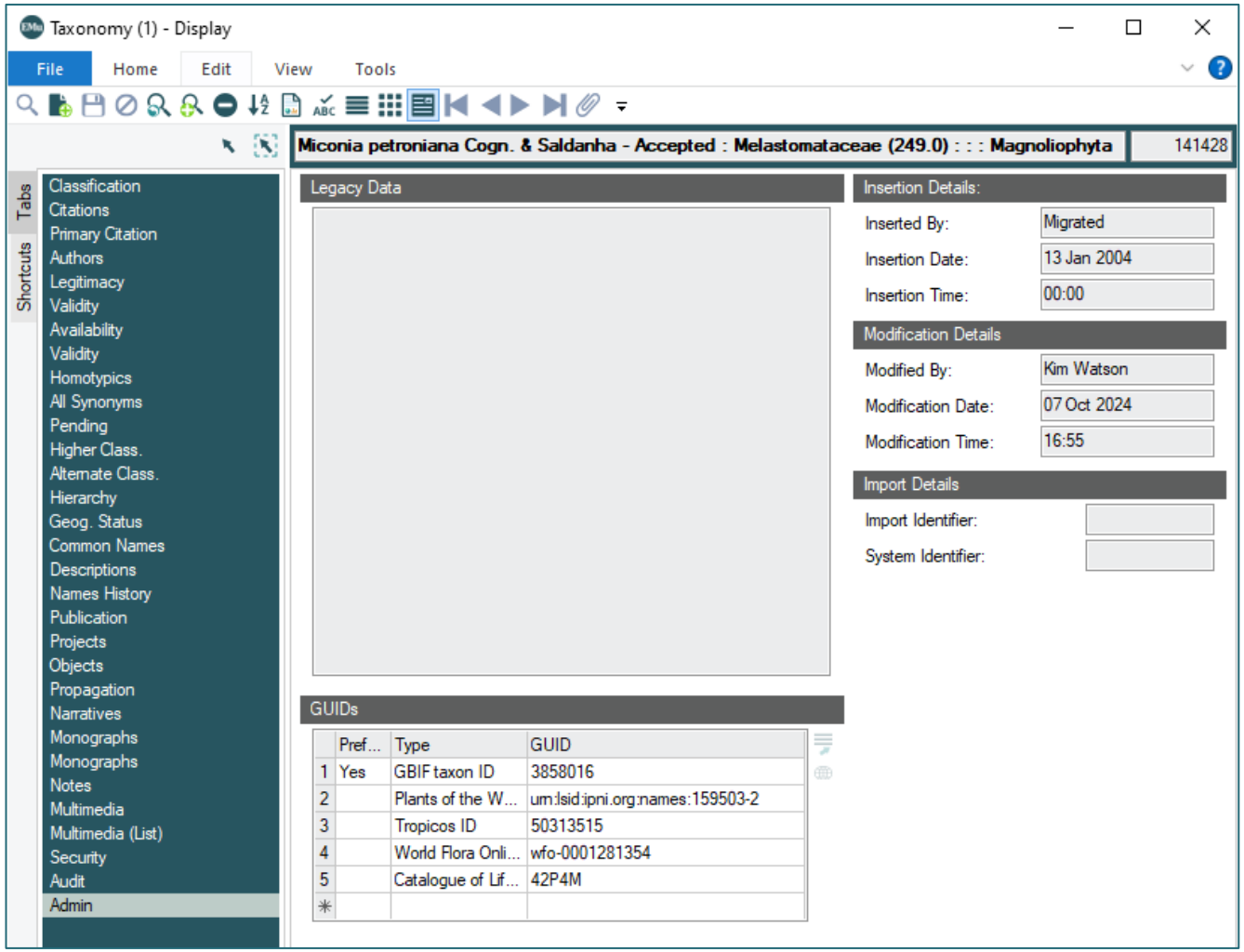
Published in: Cogn. & Saldanha ex Cogn. (1891). In: Fl. Bras. (Martius) 14(4): 612 (1888), Et DC. Monogr. Phan. 7: 834.
source: Catalogue of Life

123 OCCURRENCES

TAXON IDENTIFIERS

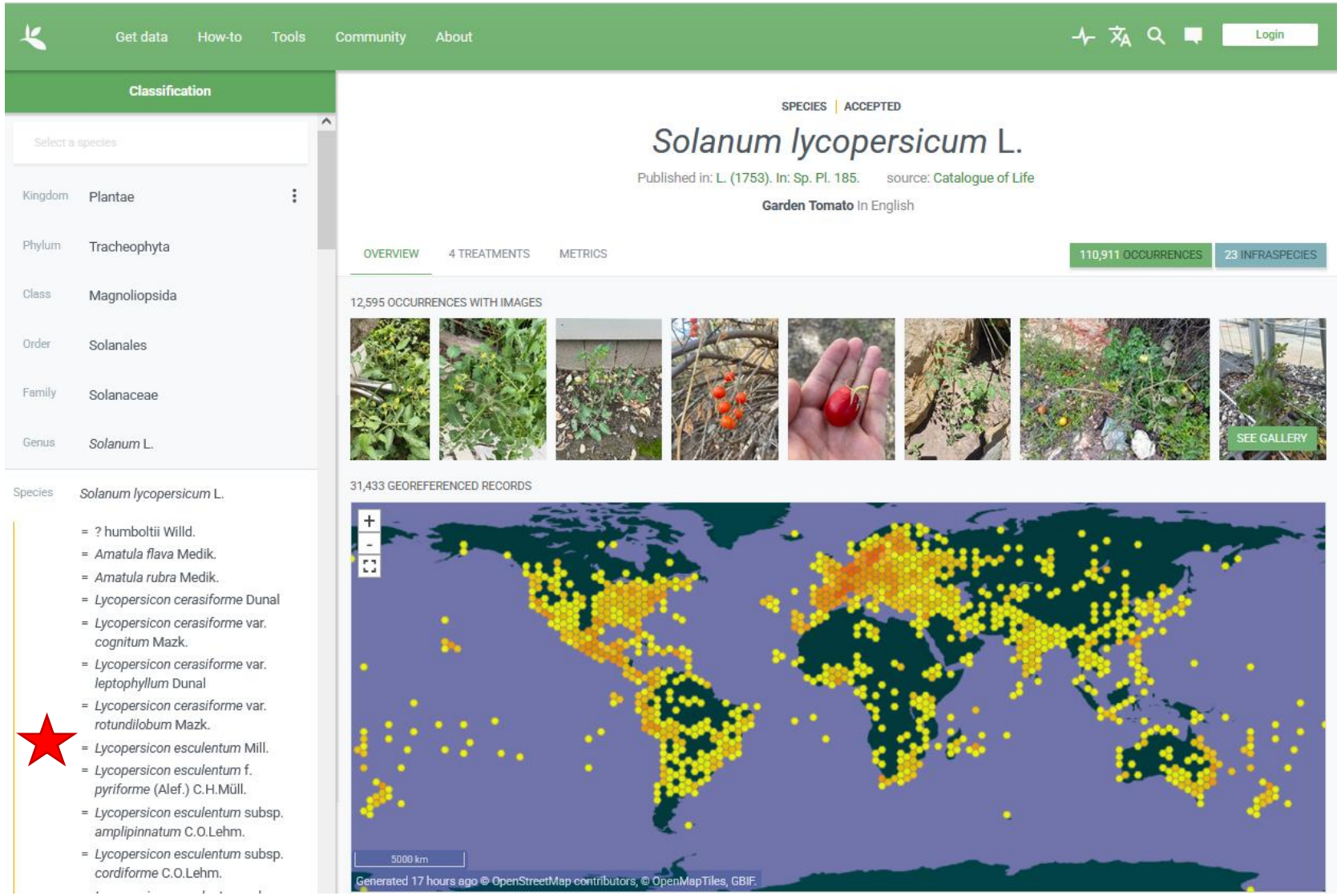
Catalogue of Life ID : 42P4M •
Plant List ID (Royal Botanic Gardens, Kew) : tro-50313515 •
IPNI plant ID : 159503-2 • Tropicos ID : 50313515 •
GBIF taxon ID : 3858016 • Encyclopedia of Life ID : 5431956 •
NCBI taxonomy ID : 2364946 •
Plants of the World Online ID : urn:lsid:ipni.org:names:159503-2 •
World Flora Online ID : wfo-0001281354 •
Open Tree of Life ID : 3941711 •
Google Knowledge Graph ID : /g/11b60v4t9v • Reflora ID : FB9745 •
UMLS CUI : C4949961 • BOLD Systems taxon ID : 1098546

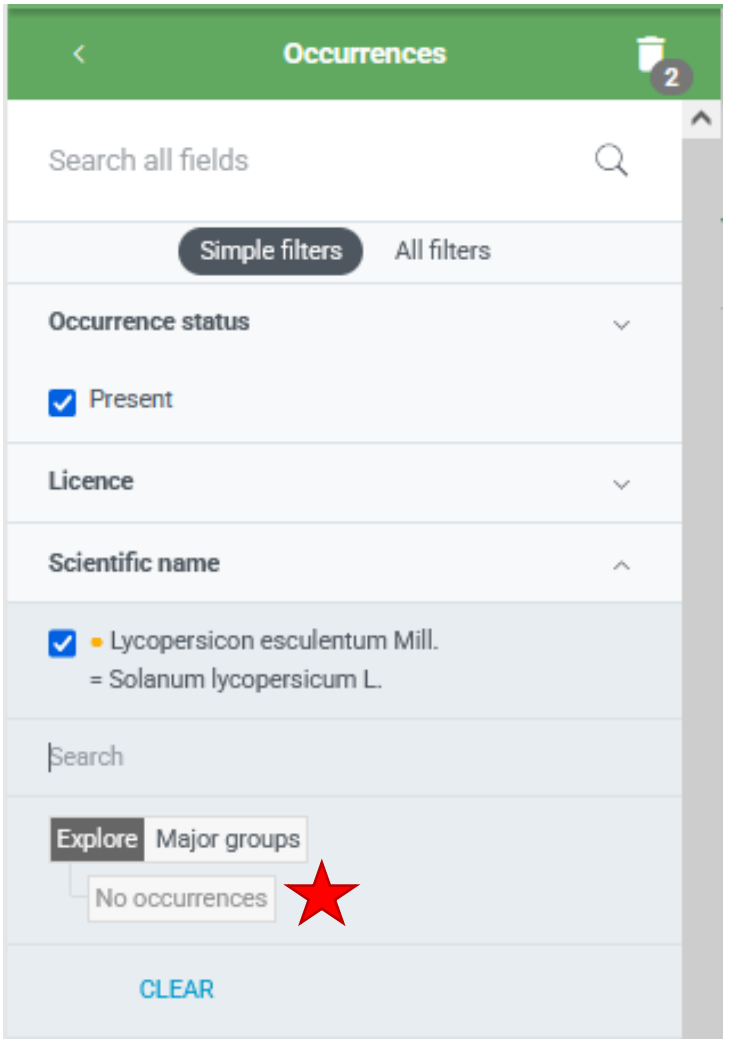
source: wikidata: Q17586449



NYBG

You say tomato… I say *Solanum lycopersicum* L.





NYBG

You say tomato… I say *Solanum lycopersicum* L.

Synced 17 days ago

OCCURRENCE | 19 JULY 2011

Lycopersicon esculentum Mill.

Divine Nightshade In English **Collected in United States of America**

Plantae › Tracheophyta › Magnoliopsida › Solanales › Solanaceae

DETAILS

GBIF Taxon interpretation: Solanum L.
Location: North America › United States of America
Basis of record: Preserved specimen
Dataset: The New York Botanical Garden Herbarium (NY)
Publisher: The New York Botanical Garden
Reference: http://sweetgum.nybg.org/science/vh/specimen_detai...
Issues: Taxon match higherrank | Modified date invalid

Taxon

Term	Interpreted	Original	Remarks
Kingdom	Plantae	Plantae	Taxon match higherrank
Phylum	Tracheophyta	Magnoliophyta	Taxon match higherrank
Class	Magnoliopsida		Taxon match higherrank
Order	Solanales	Solanales	Taxon match higherrank
Family	Solanaceae	Solanaceae	Taxon match higherrank
Genus	Solanum	Lycopersicon	Taxon match higherrank
Specific epithet		esculentum	Taxon match higherrank
Generic name	Solanum		Taxon match higherrank
Nomenclatural code	ICN	ICN	
Scientific name	Solanum L.	Lycopersicon esculentum Mill.	Taxon match higherrank
Scientific name authorship		Mill.	Taxon match higherrank
Rank	Genus	Species	Altered
Taxonomic status	Accepted		Inferred

NYBG

You say tomato… I say *Solanum lycopersicum* L.

Synced 17 days ago

OCCURRENCE | 19 JULY 2011

Lycopersicon esculentum Mill.

Divine Nightshade In English **Collected in United States of America**

Plantae › Tracheophyta › Magnoliopsida › Solanales › Solanaceae

DETAILS

GBIF Taxon interpretation: Solanum L.
Location: North America › United States of America
Basis of record: Preserved specimen
Dataset: The New York Botanical Garden Herbarium (NY)
Publisher: The New York Botanical Garden
Reference: http://sweetgum.nybg.org/science/vh/specimen_detai...
Issues: Taxon match higherrank | Modified date invalid

Taxon

Term	Interpreted	Original	Remarks
Kingdom	Plantae	Plantae	Taxon match higherrank
Phylum	Tracheophyta	Magnoliophyta	Taxon match higherrank
Class	Magnoliopsida		Taxon match higherrank
Order	Solanales	Solanales	Taxon match higherrank
Family	Solanaceae	Solanaceae	Taxon match higherrank
Genus	Solanum	Lycopersicon	Taxon match higherrank
Specific epithet		esculentum	Taxon match higherrank
Generic name	Solanum		Taxon match higherrank
Nomenclatural code	ICN	ICN	
Scientific name	Solanum L.	Lycopersicon esculentum Mill.	Taxon match higherrank
Scientific name authorship		Mill.	Taxon match higherrank
Rank	Genus	Species	Altered
Taxonomic status	Accepted		Inferred

Specimen Record

Plantae > Tracheophyta > Magnoliopsida > Solanales > Solanaceae > Solanum

Lycopersicon esculentum Mill.

From The New York Botanical Garden Herbarium (NY)

Continent	North America	Institution Code	Ny
Country	United States	Collection Code	Ny
State/	Texas	Catalog Number	02423508

Data | Flags | Raw

Taxonomy

Field	Original	Interpreted What does this mean?
Scientific Name	Lycopersicon esculentum Mill.	lycopersicon esculentum Interpreted
Kingdom	Plantae	plantae
Phylum	Magnoliophyta	tracheophyta Interpreted
Class		magnoliopsida Interpreted
Order	Solanales	solanales
Family	Solanaceae	solanaceae
Genus	Lycopersicon	solanum Interpreted
Specific Epithet	esculentum	esculentum
Scientific Name Authors	Mill.	Mill.
Taxon Rank	Species	species

NYBG

You say tomato… I say *Solanum lycopersicum* L.

Synced 17 days ago

OCCURRENCE | 19 JULY 2011

Lycopersicon esculentum Mill.

Divine Nightshade In English **Collected in United States of America**

Plantae › Tracheophyta › Magnoliopsida › Solanales › Solanaceae

DETAILS

GBIF Taxon interpretation: Solanum L.
Location: North America › United States of America
Basis of record: Preserved specimen
Dataset: The New York Botanical Garden Herbarium (NY)
Publisher: The New York Botanical Garden
Reference: http://sweetgum.nybg.org/science/vh/specimen_detai...
Issues: Taxon match higherrank | Modified date invalid

Taxon

Term	Interpreted	Original	Remarks
Kingdom	Plantae	Plantae	Taxon match higherrank
Phylum	Tracheophyta	Magnoliophyta	Taxon match higherrank
Class	Magnoliopsida		Taxon match higherrank
Order	Solanales	Solanales	Taxon match higherrank
Family	Solanaceae	Solanaceae	Taxon match higherrank
Genus	Solanum	Lycopersicon	Taxon match higherrank
Specific epithet		esculentum	Taxon match higherrank
Generic name	Solanum		Taxon match higherrank
Nomenclatural code	ICN	ICN	
Scientific name	Solanum L.	Lycopersicon esculentum Mill.	Taxon match higherrank
Scientific name authorship		Mill.	Taxon match higherrank
Rank	Genus	Species	Altered
Taxonomic status	Accepted		Inferred

Specimen Record

Plantae > Tracheophyta > Magnoliopsida > Solanales > Solanaceae > Solanum

Lycopersicon esculentum Mill.

From The New York Botanical Garden Herbarium (NY)

Continent	North America	Institution Code	Ny
Country	United States	Collection Code	Ny
State/	Texas	Catalog Number	02423508

Data | Flags | Raw

dwc_specificepithet_replaced	Darwin Core Specific Epithet (dwc:specificEpithet) replaced with a standardized value from GBIF Backbone Taxonomy.
dwc_taxonomicstatus_added	Darwin Core Taxonomic Status (dwc:taxonomicStatus) added where none was provided.
gbif_genericname_added	GBIF Generic Name added from GBIF Backbone Taxonomy.
dwc_datasetid_added	Darwin Core Dataset ID (dwc:datasetID) added where none was provided.
gbif_taxon_corrected	A match in GBIF Backbone Taxonomy was found. Inverse of taxon_match_failed flag.
dwc_scientificnameauthorship_replaced	

NYBG

Controlled Vocabulary

Lookup List (Pick List) = List of values that have been used or are suggested for use in a field

Benefits of controlled vocabulary:

- Efficiency
- Consistency
- Use of correct or approved terminology
- Decreased heterogeneity = Increased searchability = Increased usability

Lookup Lists may be:

- Pre-populated, editable, read-only (select existing value)
- Linked together in a logical, hierarchical relationship, which allows auto-filling and filtering of values (e.g., taxonomy, geography)

Note: Database may/may not support Unicode, so values including/not punctuation and diacritics are significant

Darwin Core terms that recommend Controlled Vocab

Biodiversity Information Standards TDWG

	Term	Recommended
Record-level	dcterms:type	DCMI Type Vocabulary
	dcterms:language	RFC 4646
	basisOfRecord	Darwin Core classes
Occurrence	sex	
	lifeStage	
	reproductiveCondition	
	behavior	
	establishmentMeans	
	occurrenceStatus	
	disposition	
Organism	organismScope	
Event	sampleSizeUnit	Ontology of Units of Measure
Location	higherGeographyID	Getty Thesaurus of Geographic Names
	continent	Getty Thesaurus of Geographic Names
	waterbody	Getty Thesaurus of Geographic Names
	islandGroup	Getty Thesaurus of Geographic Names
	island	Getty Thesaurus of Geographic Names
	country	Getty Thesaurus of Geographic Names
	countryCode	ISO 3166-1-alpha-2
	geodeticDatum	EPSG
	verbatimCoordinateSystem	
	verbatimSRS	EPSG
	georeferenceVerificationStatus	{'requires verification', 'verified by collector', or 'verified by curator' }

https://github.com/tdwg/dwc-qa/wiki/Controlled-Vocabularies

NYBG

Darwin Core terms that recommend Controlled Vocab

Biodiversity Information Standards TDWG

	Term	Recommended
Identification	identificationVerificationStatus	HISPID/ABCD
Taxon	taxonRank	
	nomenclaturalCode	
	taxonomicStatus	
MeasurementOrFact	measurementType	
	measurementUnit	International System of Units (SI)
ResourceRelationship	relationshipOfResource	

DwC Preliminary compilation of controlled vocabularies: ControlledVocabs Resources.

GBIF's Vocabulary Server: Tool to standardize selected fields

https://github.com/tdwg/dwc-qa/wiki/Controlled-Vocabularies

Svenningsen, Cecilie & Gonzalez, Marcos & Robertson, Tim. (2024). GBIF's Vocabulary Server: A Tool to Create, Manage and Apply Controlled Vocabularies for Biodiversity. *Biodiversity Information Science and Standards*. 8. e137853. https://doi.org/10.3897/biss.8.137853

NYBG

Darwin Core terms with inherent restrictions

Biodiversity Information Standards TDWG

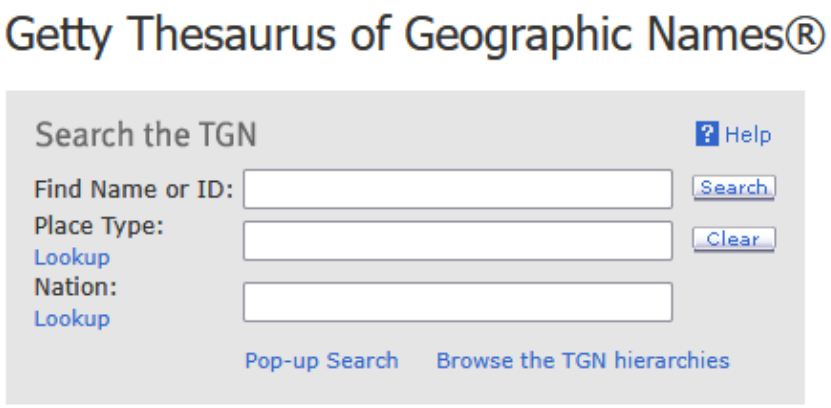
	Term	Restriction
Record-level	dcterms:modified	ISO 8601:2004(E) <= now
Occurrence	individualCount	positive integer or 0
Event	eventDate	ISO 8601:2004(E) <= now
	eventTime	ISO 8601:2004(E) <= now
	startDayOfYear	positive integer <= 366 (or 365)
	endDayOfYear	positive integer <= 366 (or 365)
	year	integer <= current year
	month	positive integer <= 12
	day	positive integer <= 31 (or 30, or 28)
Location	decimalLatitude	real number between -90 and 90 inclusive
	decimalLongitude	real number between -180 and 180 inclusive
	coordinateUncertaintyInMeters	real number > 0
	coordinatePrecision	subset of positive real numbers> 0
	pointRadiusSpatialFit	0 or positive real number >= 1
	footprintWKT	valid geometry in Well-known Text
	footprintSRS	valid SRS in Well-known Text
	footprintSpatialFit	0 or positive real number >= 1
	georeferencedDate	ISO 8601:2004(E) <= now
Identification	dateIdentified	ISO 8601:2004(E) <= now
Taxon	namePublishedInYear	four-digit year
MeasurementOrFact	measurementDeterminedDate	ISO 8601:2004(E) <= now
ResourceRelationship	relationshipEstablishedDate	ISO 8601:2004(E) <= now

https://github.com/tdwg/dwc-qa/wiki/Controlled-Vocabularies

NYBG

Controlled Vocabulary: Geography

	Term	Recommended
Location	higherGeographyID	Getty Thesaurus of Geographic Names
	continent	Getty Thesaurus of Geographic Names
	waterbody	Getty Thesaurus of Geographic Names
	islandGroup	Getty Thesaurus of Geographic Names
	island	Getty Thesaurus of Geographic Names
	country	Getty Thesaurus of Geographic Names
	countryCode	ISO 3166-1-alpha-2
	geodeticDatum	EPSG
	verbatimCoordinateSystem	
	verbatimSRS	EPSG
	georeferenceVerificationStatus	{'requires verification', 'verified by collector', or 'verified by curator' }



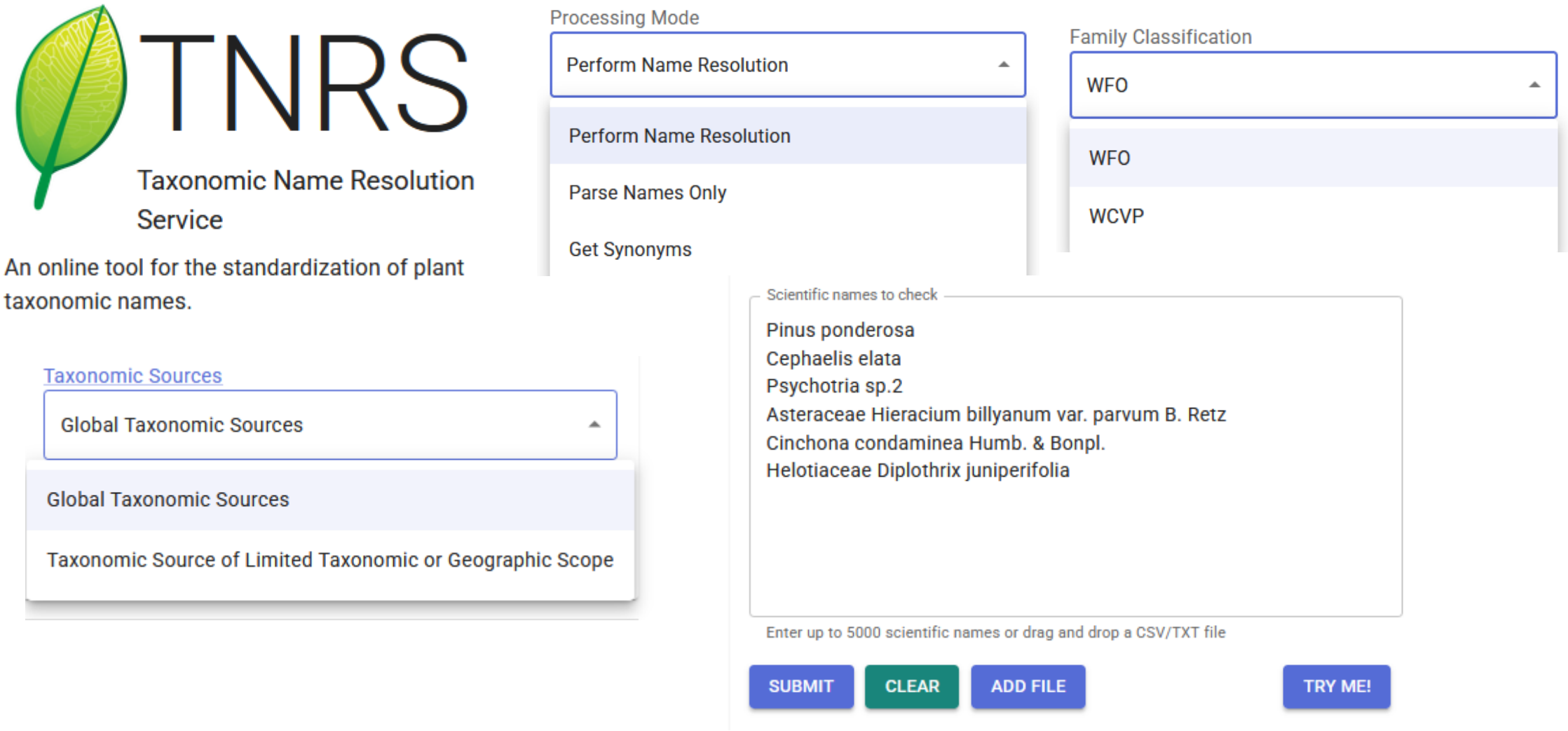
- **>3m place records (>5m names)**
- **A thesaurus, not a GIS**
- **Focus on names, historical information; is hierarchical**
- **Multilingual; not comprehensive**
- **Coordinates are approximate**

https://www.getty.edu/research/tools/vocabularies/tgn/

NYBG

Data Cleanup

Data Cleanup: Taxonomy



Boyle, B. L., Matasci, N., Mozzherin, D., Rees, T., Barbosa, G. C., Kumar Sajja, R., & Enquist, B. J. (2021). Taxonomic Name Resolution Service, version 5.1. In *Botanical Information and Ecology Network*. https://tnrs.biendata.org/

Data Cleanup: Taxonomy

TNRS

Taxonomic Name Resolution Service

An online tool for the standardization of plant taxonomic names.

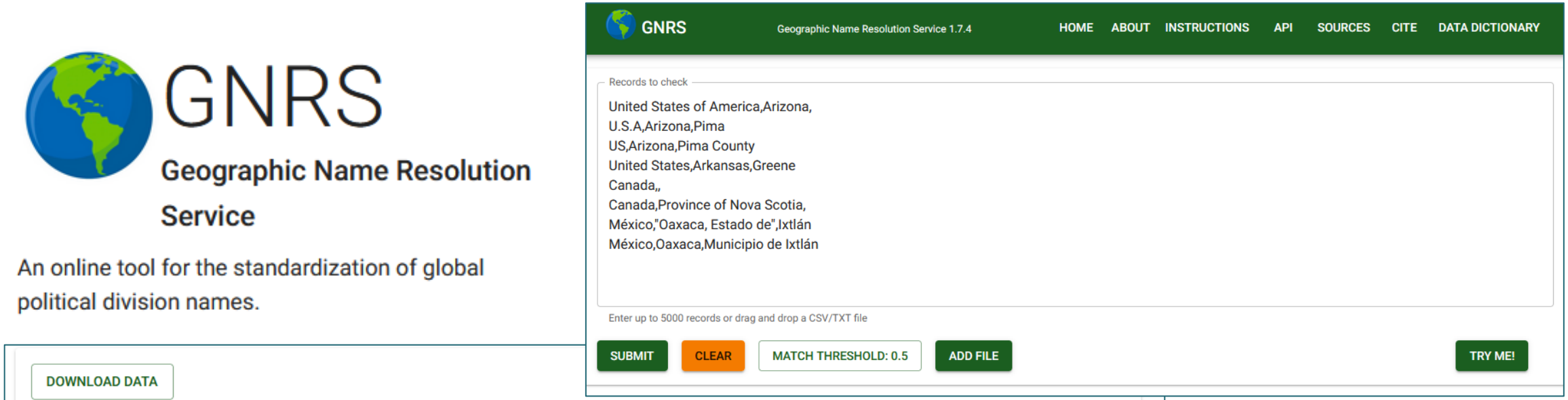
BEST MATCH SETTINGS | MATCH 0.53 | DOWNLOAD DATA | DOWNLOAD SETTINGS

Warnings	Name Submitted	Name Matched	Source	Overall Score	Unmatched Terms	Taxonomic Status	Accepted Name	Details
	Pinus ponderosa	Pinus ponderosa Douglas ex. C.Lawson (+1 more)	wfo	1.00		Accepted	Pinus ponderosa Douglas ex. C.Lawson	Details
	Cephaelis elata	Cephaelis elata Sw.	wfo,wcvp	1.00		Synonym	Palicourea elata (Sw.) Borhidi	Details
	Cinchona condaminea Humb. & Bonpl.	Cinchona condaminea Bonpl.	wfo,wcvp	.92		Synonym	Cinchona officinalis L.	Details
	Asteraceae Hieracium billyanum var. parvum B. Retz	Hieracium billyanum var. parvulum de Retz (+1 more)	wfo	.90		Synonym	Pilosella billyana (de Retz) Mateo	Details
	Psychotria sp.2	Psychotria L.	wfo,wcvp	.90	sp.2	Accepted	Psychotria L.	Details
⚠	Helotiaceae Diplothrix juniperifolia	Diplothrix juniperifolia DC. (+1 more)	wfo,wcvp	.60	Helotiaceae	Synonym	Zinnia juniperifolia A.Gray	Details

Boyle, B. L., Matasci, N., Mozzherin, D., Rees, T., Barbosa, G. C., Kumar Sajja, R., & Enquist, B. J. (2021). Taxonomic Name Resolution Service, version 5.1. In *Botanical Information and Ecology Network*. https://tnrs.biendata.org/

NYBG

Data Cleanup: Geography



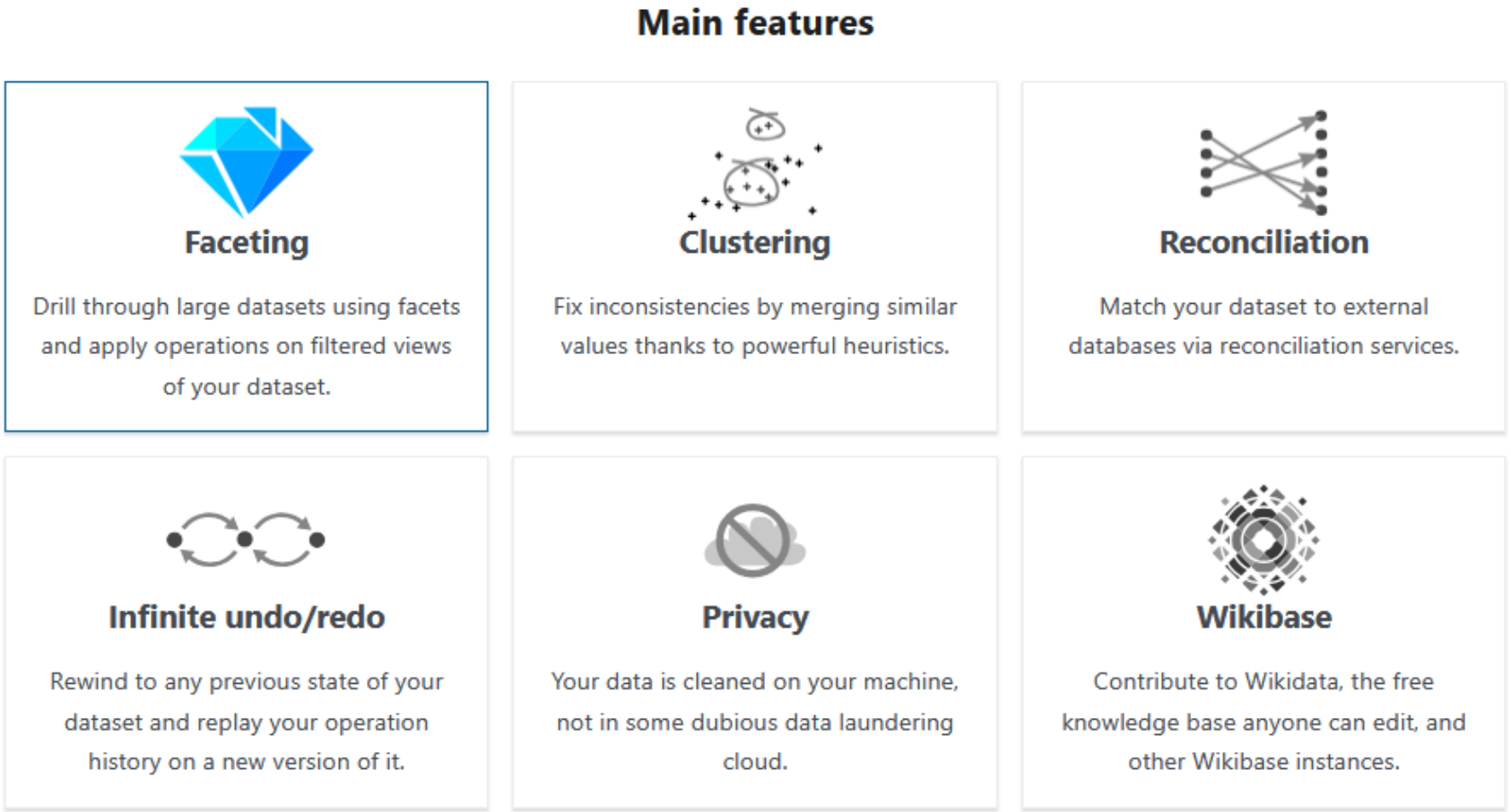
Political Division Submitted	Country	State Province	County	Overall Score	Details
Canada	Canada			1.00	Details
Canada:Province of Nova Scotia	Canada	Nova Scotia		1.00	Details
México:Oaxaca, Estado de:Ixtlán	Mexico	Oaxaca	Ixtlan de Juarez	1.00	Details
México:Oaxaca:Municipio de Ixtlán	Mexico	Oaxaca	Ixtlan de Juarez	1.00	Details
U.S.A:Arizona:Pima	United States	Arizona	Pima	1.00	Details
US:Arizona:Pima County	United States	Arizona	Pima	1.00	Details
United States of America:Arizona	United States	Arizona		1.00	Details
United States:Arkansas:Greene	United States	Arkansas	Greene	1.00	Details

Boyle, B. L., Maitner, B., Barbosa, G. C., Rethvick, S. Y. B., & Enquist, B. J. (2024). Geographic Name Resolution Service. In *Botanical Information and Ecology Network*. https://gnrs.biendata.org/

NYBG

Data Cleanup: OpenRefine

"A powerful free, open source tool for working with messy data: cleaning it, transforming it from one format into another, and extending it with web services and external data."



https://openrefine.org/

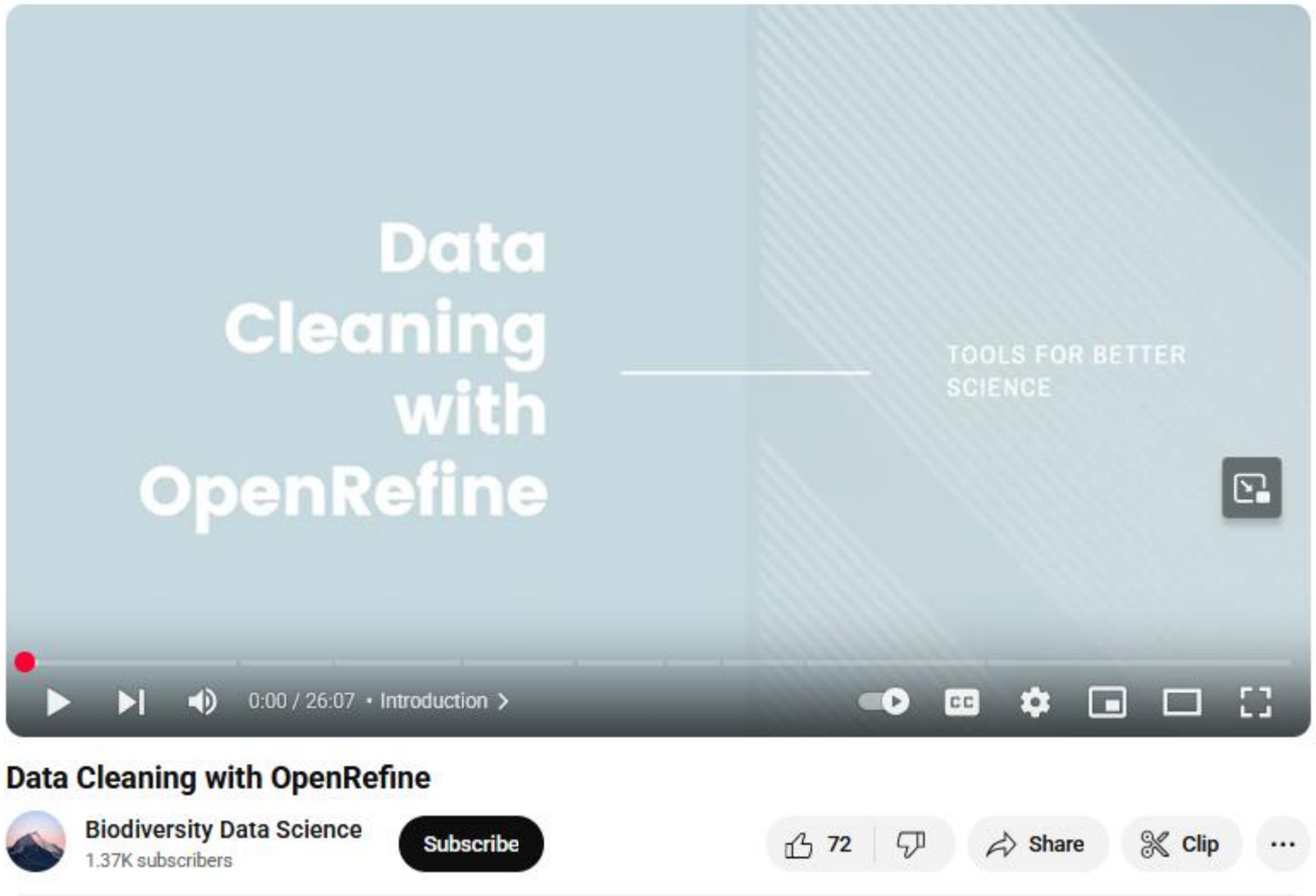
NYBG

Data Cleanup: OpenRefine

Data Cleaning with OpenRefine

Helpful tutorial video published by Biodiversity Data Science using biodiversity data

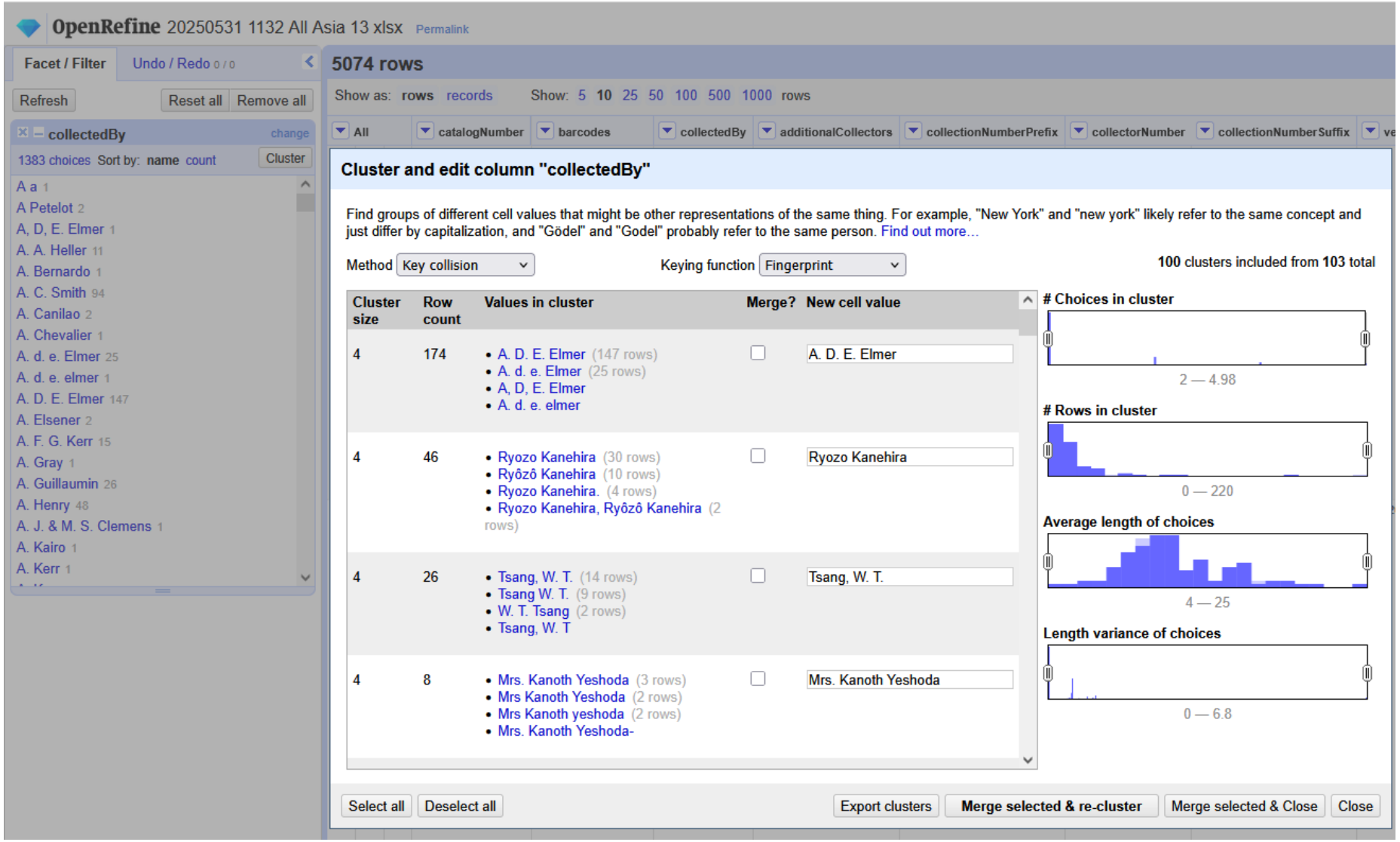
- Create Project
- Faceting
- Clusters
- Splitting
- Undo/Redo
- Trim White Spaces
- Filter/Sort
- Scripts



https://youtu.be/jyUlT8ohlG4?si=SANS46hHXu6ah1sx

NYBG

Data Cleanup: OpenRefine Demo



https://openrefine.org/

NYBG

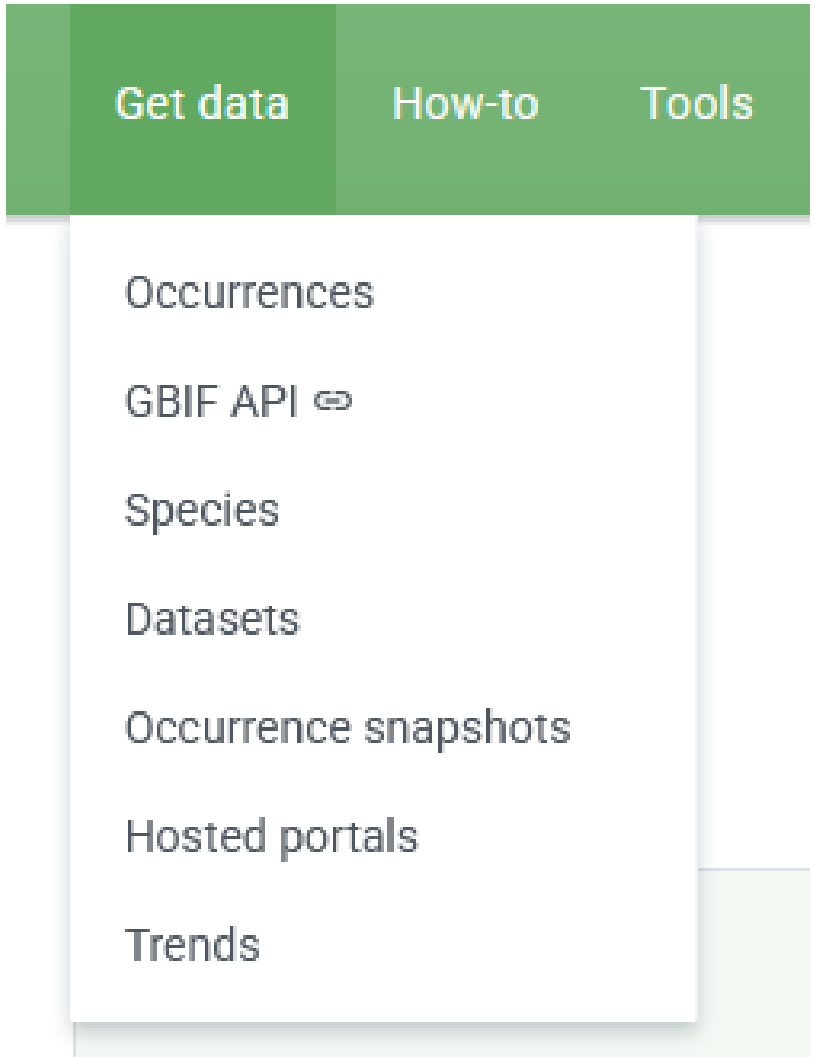
Data Aggregators

Data Aggregators: GBIF

Global Biodiversity Information Facility (GBIF)

"An international network and data infrastructure funded by the world's governments and aimed at providing anyone, anywhere, open access to data about all types of life on Earth."

Sweet woodruff (*Galium odoratum*) observed near Vestland, Norway. Photo by Jenny Smedmark via University of Bergen (CC BY 4.0)

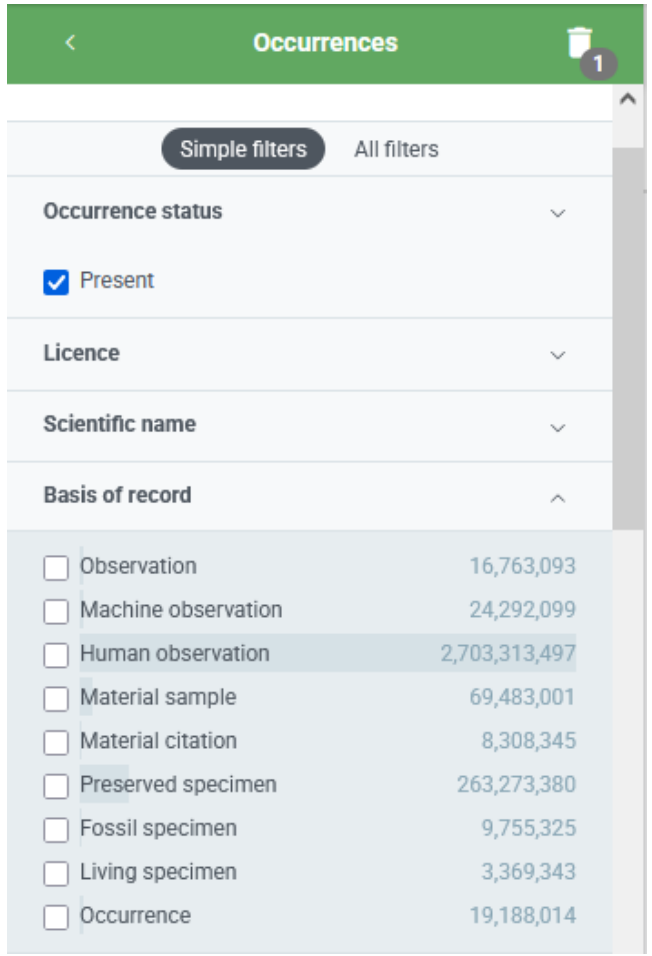


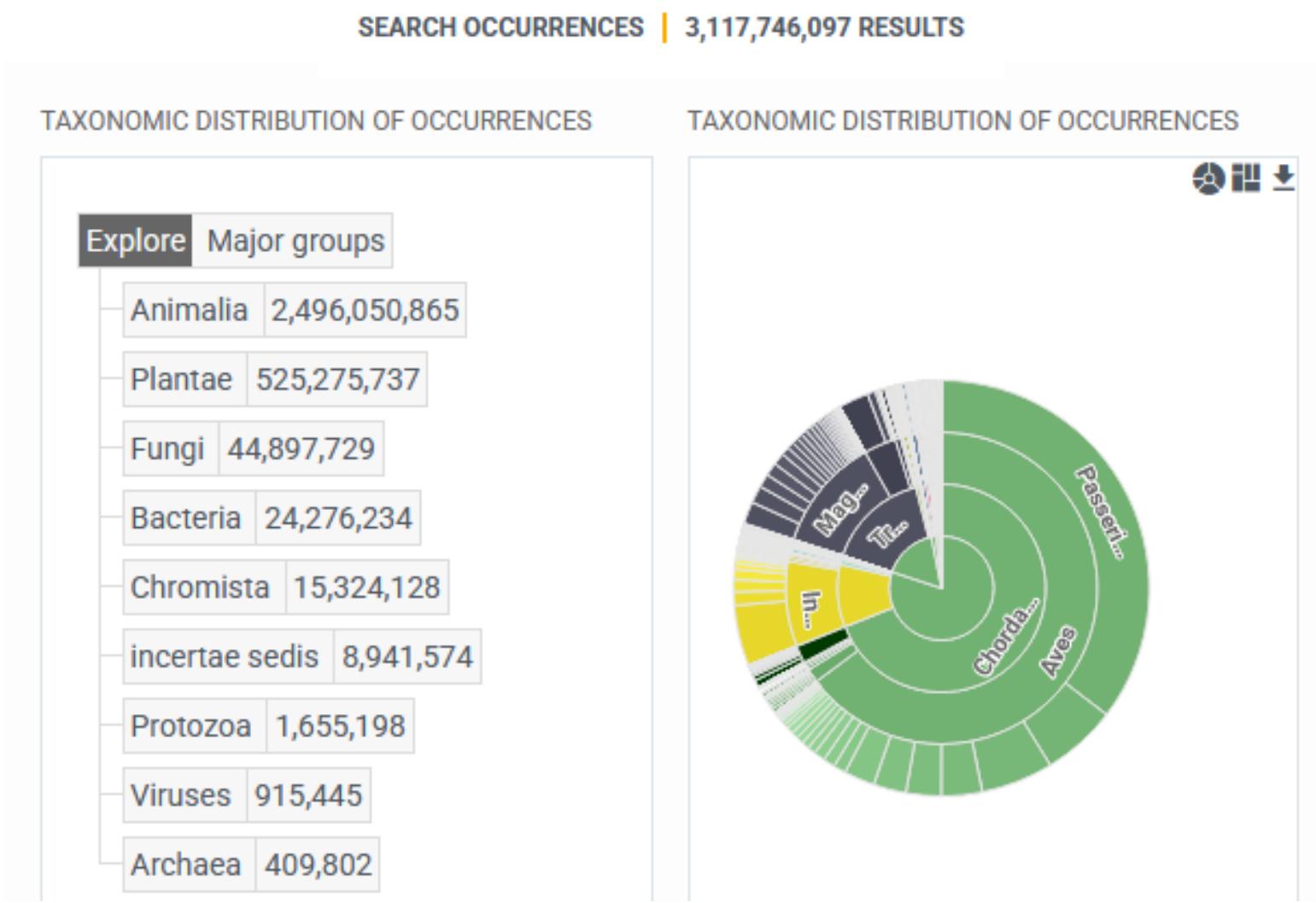
NYBG

Data Aggregators: GBIF

GBIF Global Biodiversity Information Facility

Simple Filters or All Filters

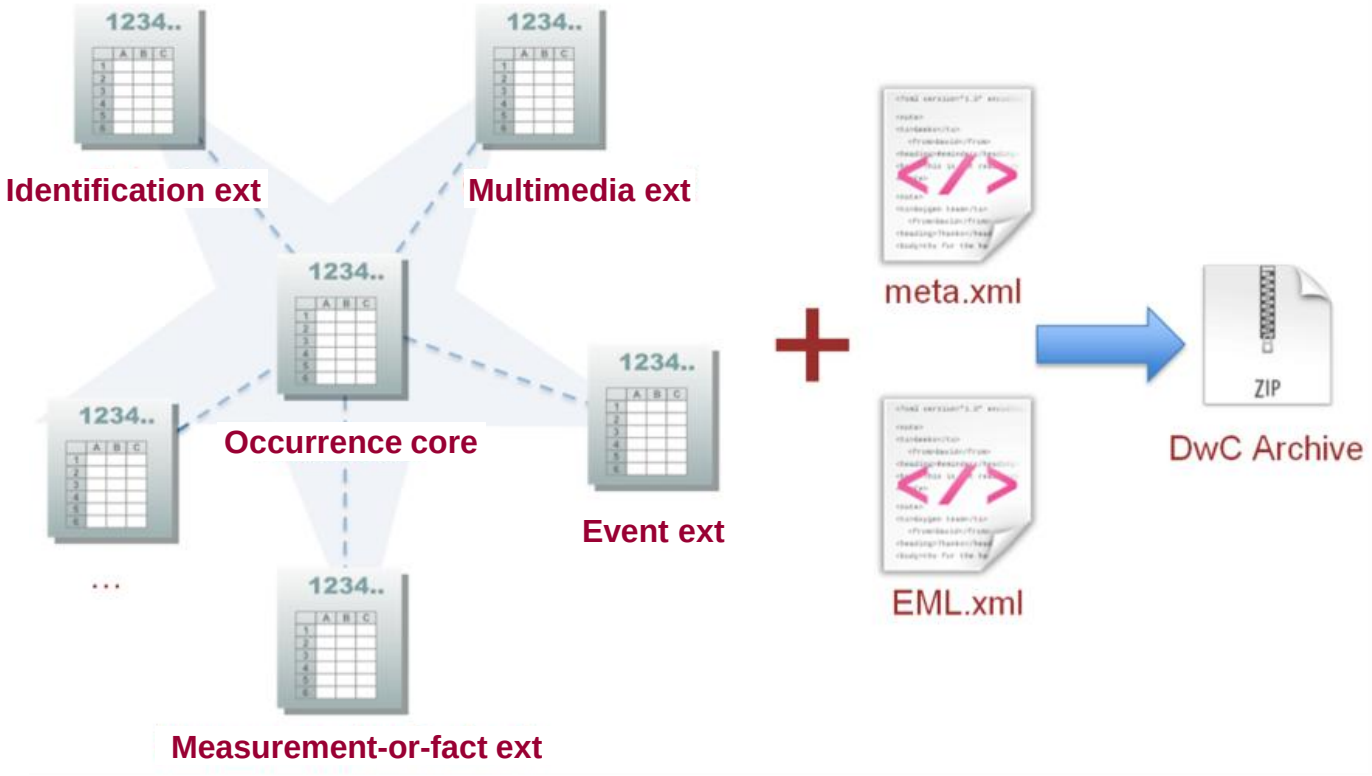




https://www.gbif.org/occurrence/download?occurrence_status=present

NYBG

Data Aggregators: GBIF



Darwin Core Archive (DwC-A)

- **Core table** with species occurrences:
  - occurrenceID = Catalogue GUID
  - Event, Location, Taxon, Identification in 1 table (no IDs).
- **Multimedia Extension table:**
  - Audiovisual Core standard
  - multimediaID = Multimedia GUID
- Link core & extension with occurrenceID

https://www.gbif.org/

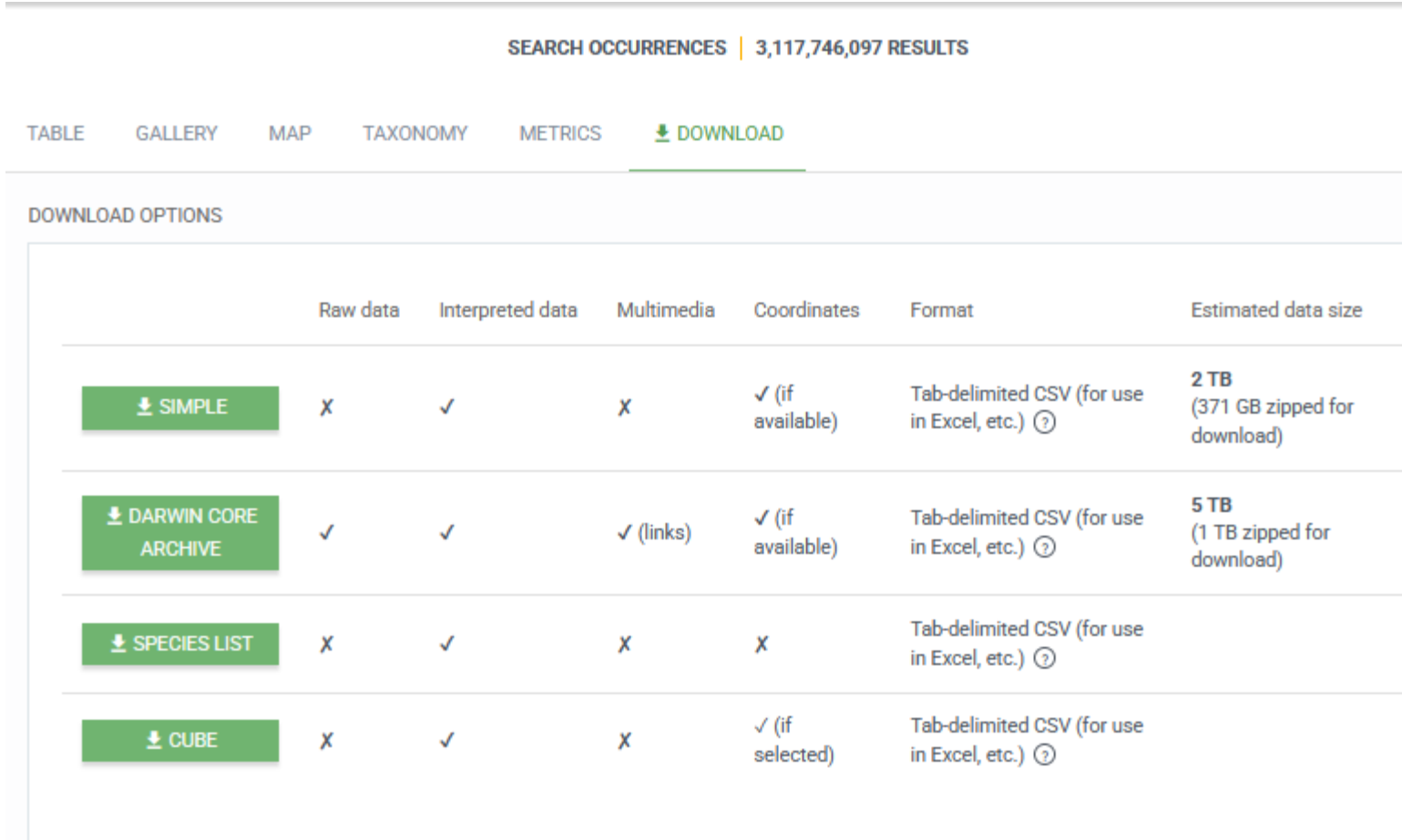
NYBG

Data Aggregators: GBIF

GBIF Global Biodiversity Information Facility

Download datasets

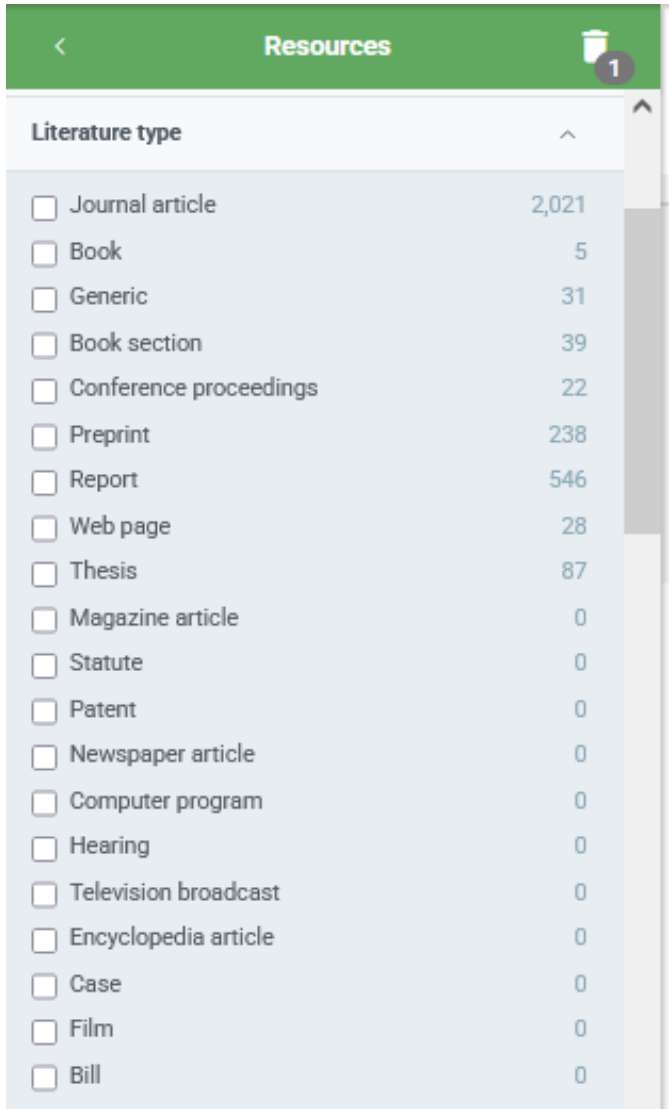
- Include records from many publishers
- Unique DOI assigned to dataset
- Abide by GBIF User Agreement
- If use the data, cite dataset appropriately, and include the unique DOI

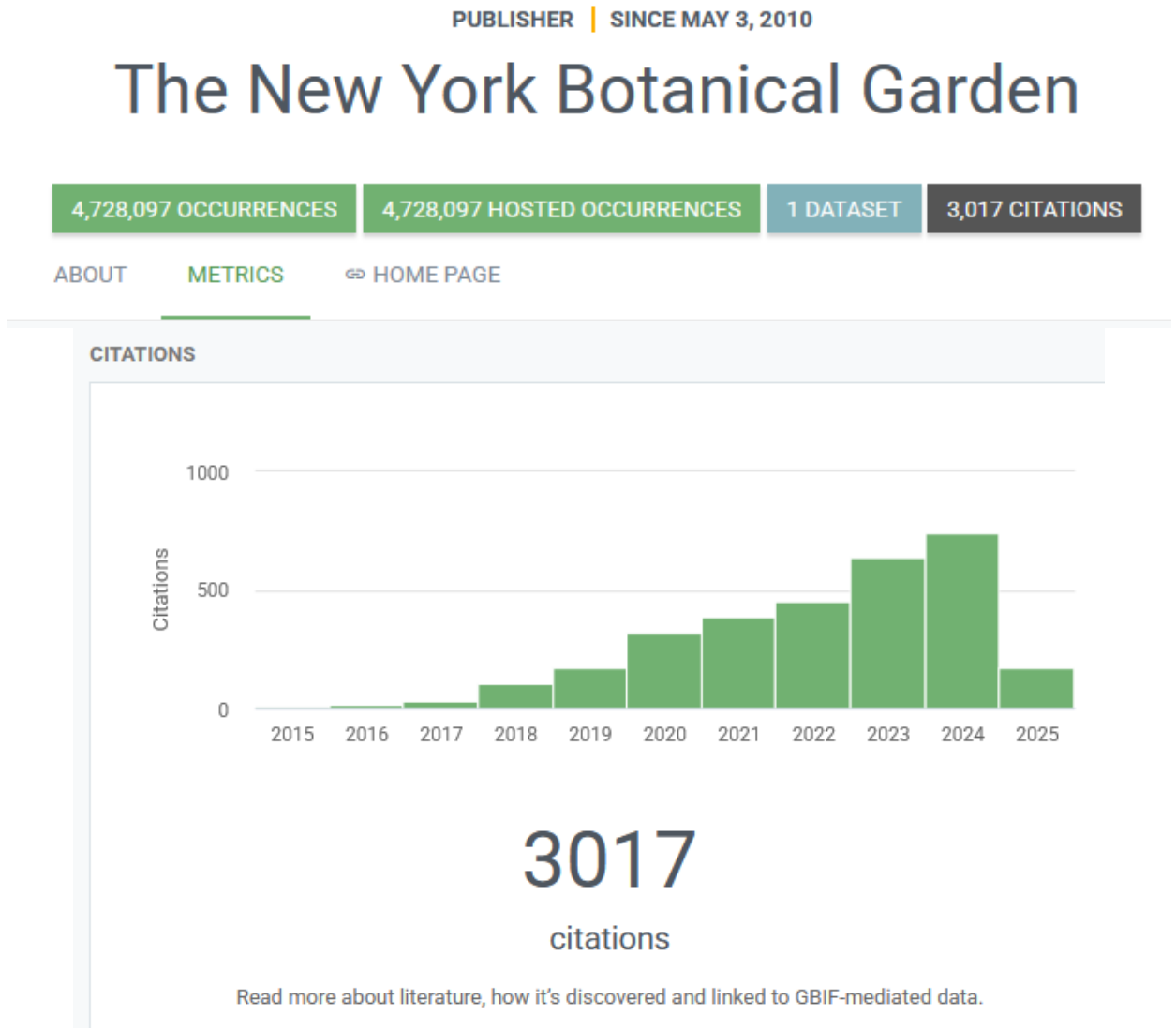


https://www.gbif.org/

NYBG

Data Aggregators: GBIF

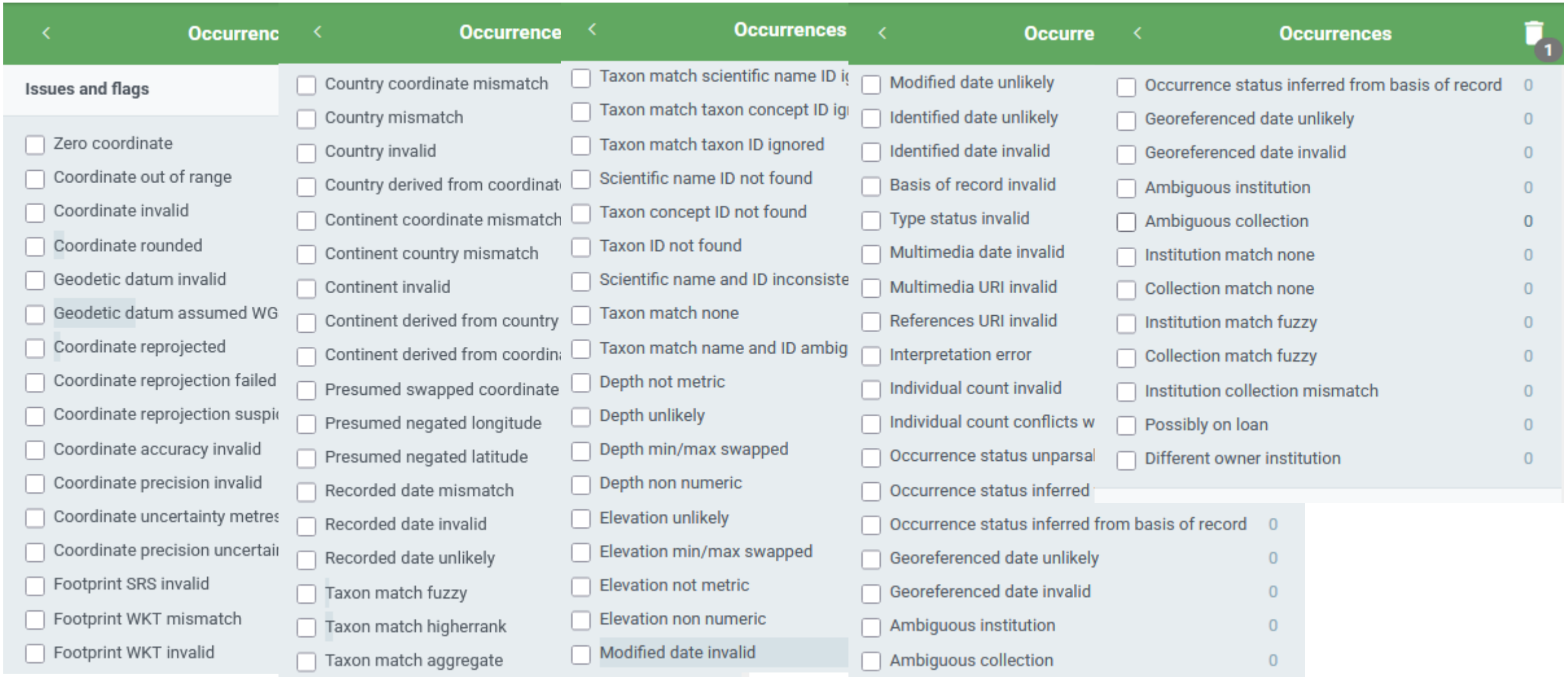




https://www.gbif.org/resource/search?contentType=literature&publishingOrganizationKey=ae447c50-b8a8-11d8-92a4-b8a03c50a862

Data Aggregators: GBIF

GBIF Global Biodiversity Information Facility



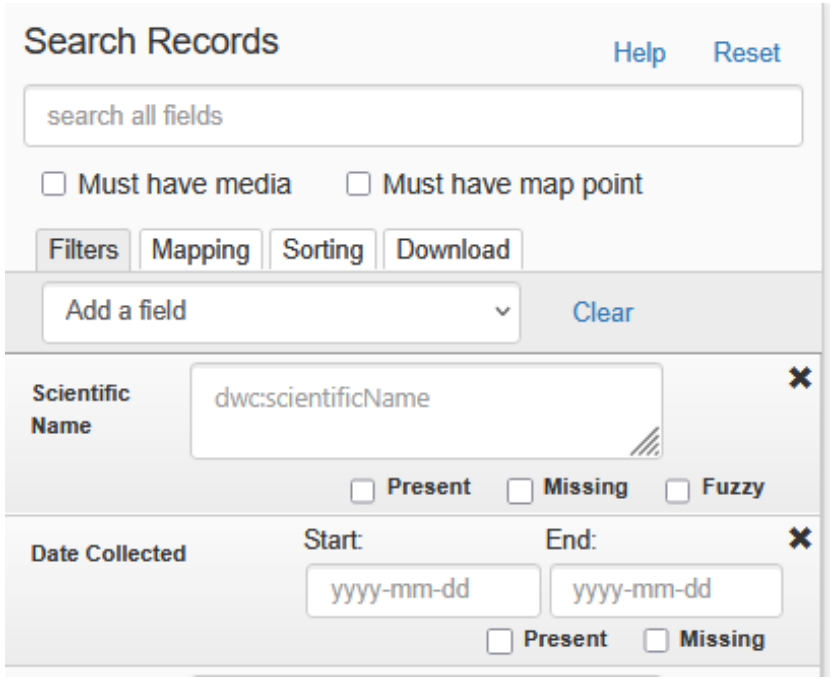
Issues and Flags for https://www.gbif.org/occurrence/search?publishing_org=ae447c50-b8a8-11d8-92a4-b8a03c50a862

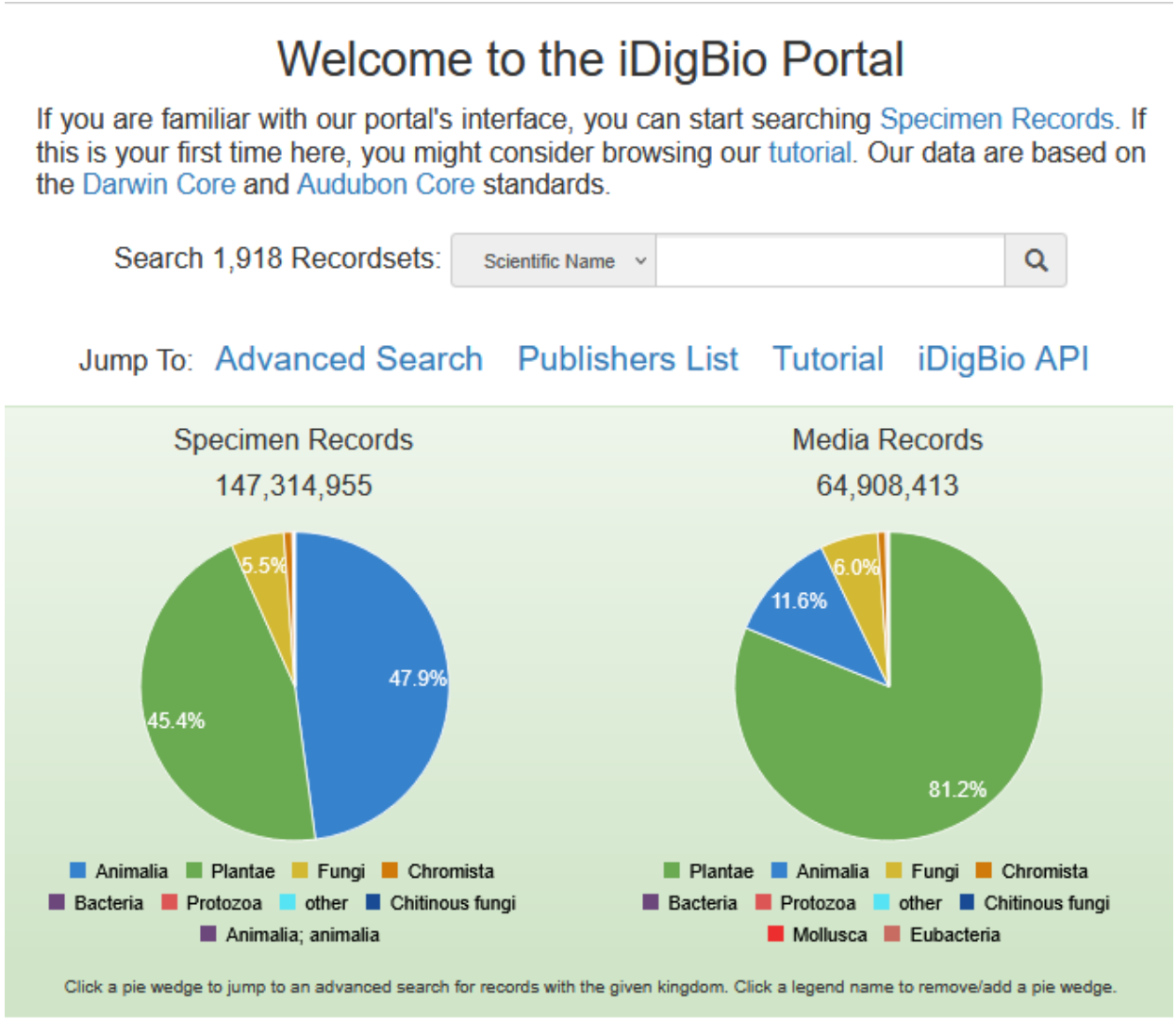
NYBG

Data Aggregators: iDigBio

iDigBio: "Mission to promote and catalyze digitization, mobilization, and use of data about biodiversity specimens through training, open data, and innovative uses of these data."

Apply filters to narrow results





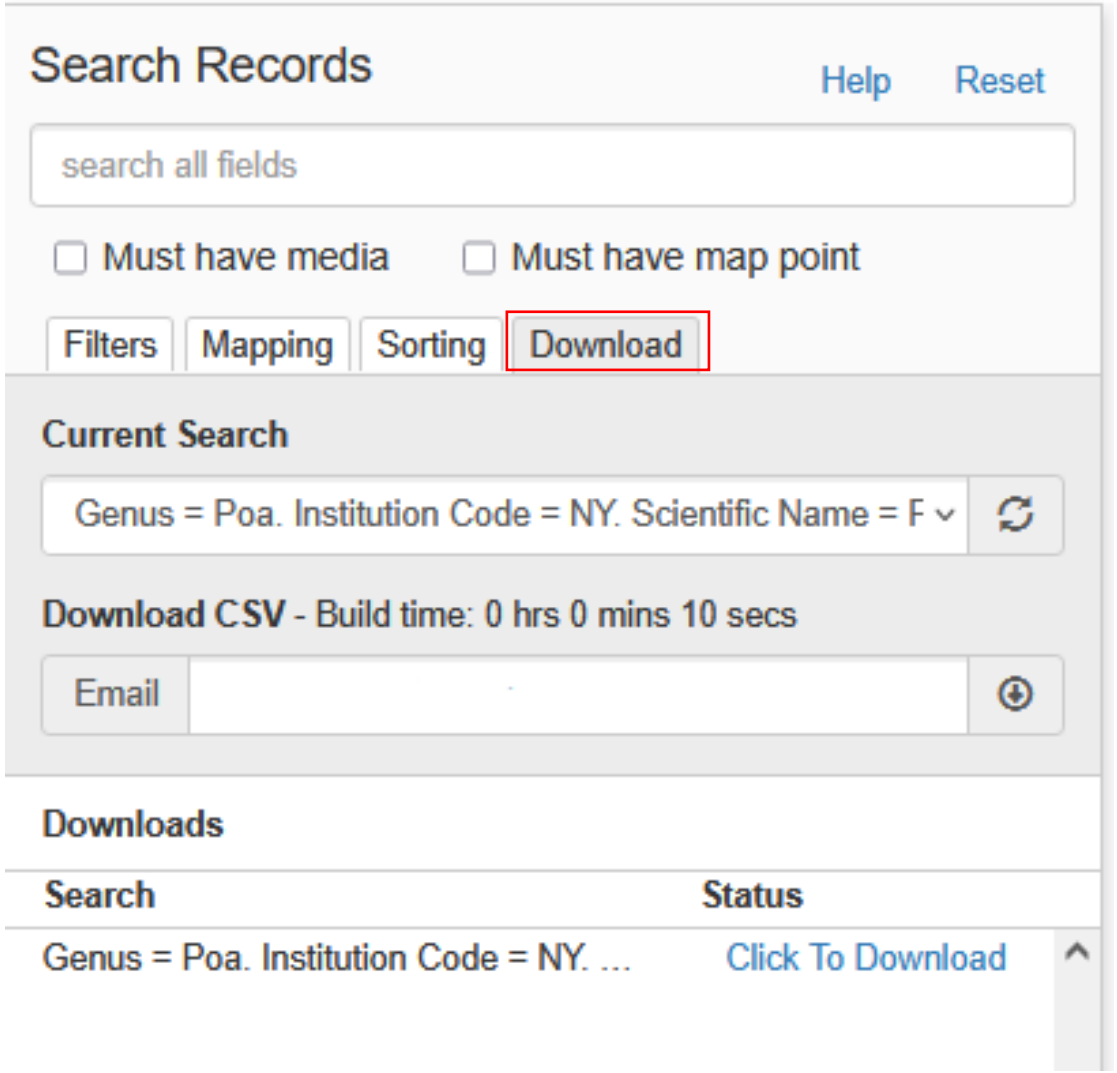
https://portal.idigbio.org/portal/

NYBG

Data Aggregators: iDigBio

Download datasets

- Include records from many publishers
- No DOI generated
- If use the data, recommend using [GBIF's Citation-Guidelines]
- Downloaded zip folder contains file **records.citation.txt** that follows GBIF's guidelines



https://portal.idigbio.org/portal/; https://www.idigbio.org/sites/default/files/iDigBio_Terms_of_Use_20180315.pdf

NYBG

Data Aggregators: iDigBio

Recordset

Search Recordset

The New York Botanical Garden Herbarium (NY)

Specimen Records: 4,727,195
Media Records: 4,057,580
iDigBio Last Ingested Date: 2025-05-23

A complete list of the data quality flags and their descriptions can be found here.

Data Corrected | Data Use | Raw

This table shows any data corrections that were performed on this recordset to improve the capabilities of iDigBio Search. The first column represents the correction performed. The last two columns represent the number and percentage of records that were corrected. A complete list of the data quality flags and their descriptions can be found here. Clicking on a data flag name will take you to a search for all records with this flag in this recordset.

Flag	Records With This Flag	(%) Percent With This Flag
dwc_datasetid_added	4133963	87.451
dwc_parentnameusageid_added	4133963	87.451
dwc_taxonid_added	4133963	87.451
dwc_taxonomicstatus_added	4133963	87.451
gbif_canonicalname_added	4133963	87.451
gbif_genericname_added	4133963	87.451
gbif_taxon_corrected	4133963	87.451
dwc_class_added	3723037	78.758
gbif_reference_added	3323544	70.307
idigbio_isocountrycode_added	3300196	69.813
dwc_scientificnameauthorship_replaced	3147084	66.574
gbif_vernacularname_added	2873682	60.79
dwc_phylum_replaced	2848679	60.262
dwc_multimedia_added	2332284	49.338

https://portal.idigbio.org/portal/recordsets/36d35b23-113e-4633-90ec-19d265a3b5f6

Data Aggregators: Symbiota Portals

Distributed network of theme-based research portals

60+ portals, 95m occurrences, 43m images

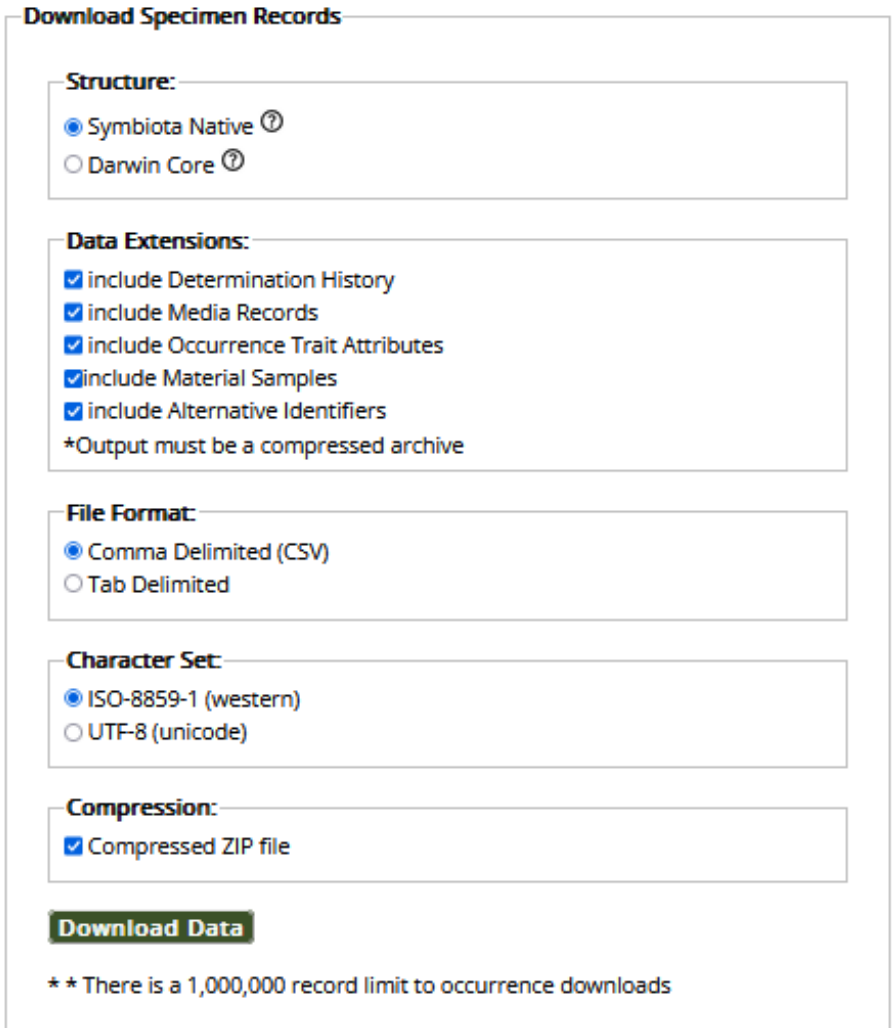
Search terms, Lat/Long, filter by Collections

https://symbiota.org/

Data Aggregators: Symbiota Portals

Download datasets

- Include records from many publishers
- Abide by data usage terms for each portal
- 1m record limit to occurrence downloads
- If use the data, recommend using [GBIF's Citation-Guidelines]



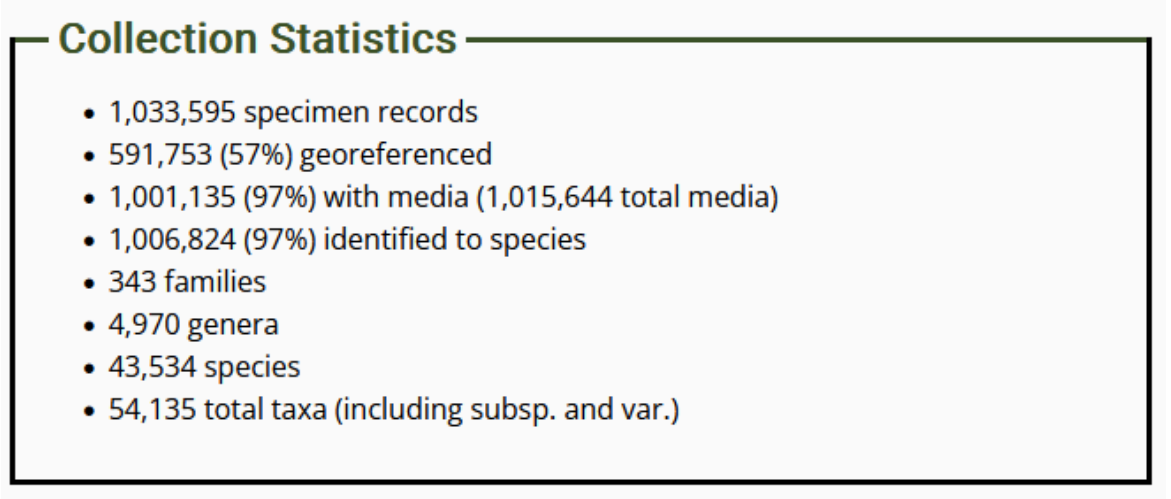
Data Aggregators: Symbiota Portals

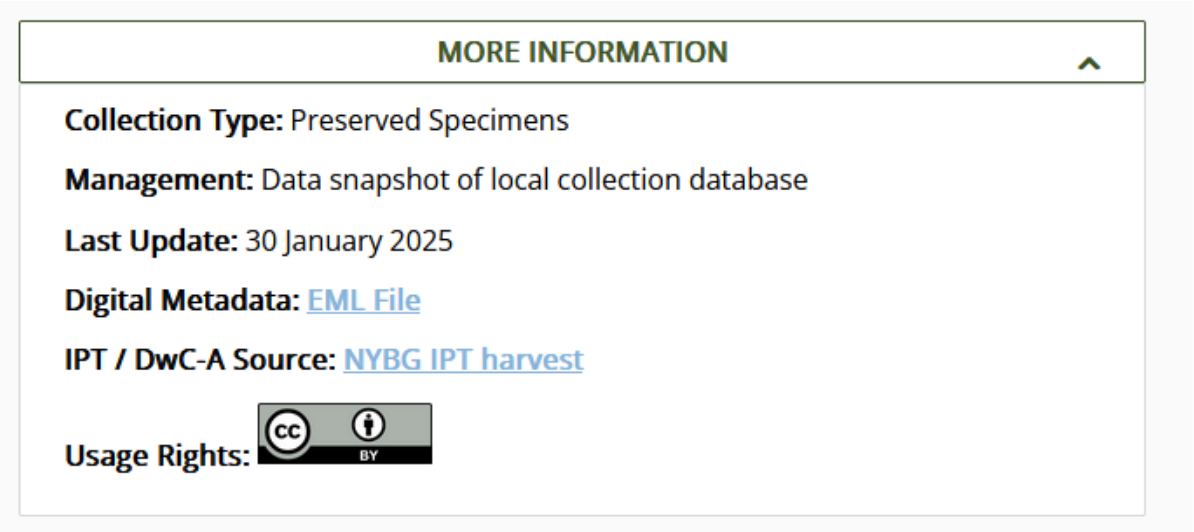
Collections Profile includes

- Contacts
- Collection Statistics
- Data flags?

Must be Admin user on each portal to:

- Edit profile
- Review posted comments





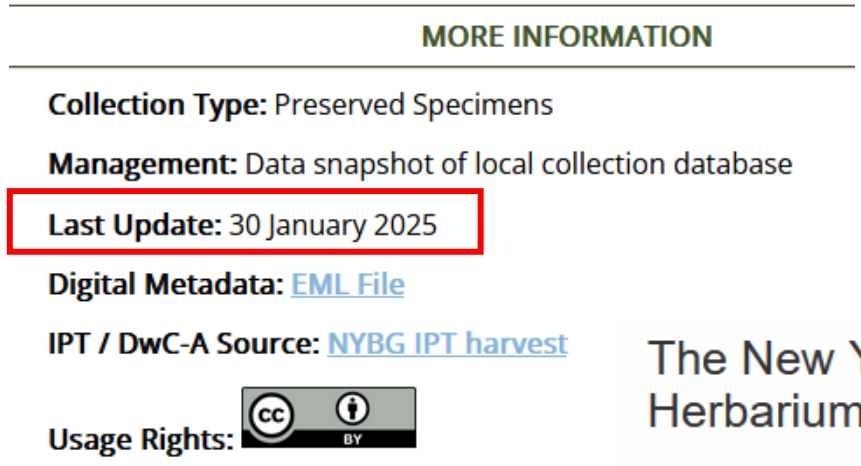
Data Synchrony & Feedback Loop

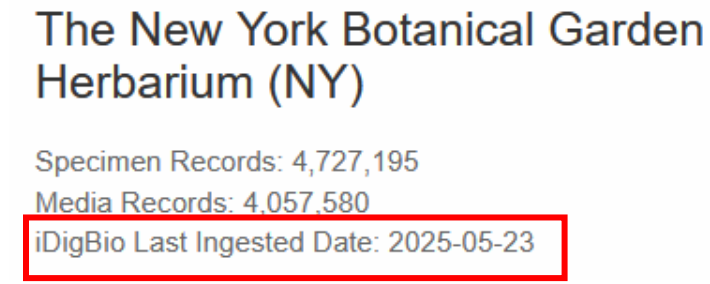
Data Synchrony:

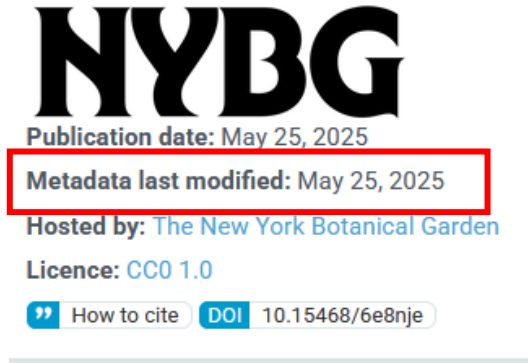
- **CAUTION:** Aggregators refresh data snapshots on variable schedule(s)
- = Data are out of sync with recent edits made to source database

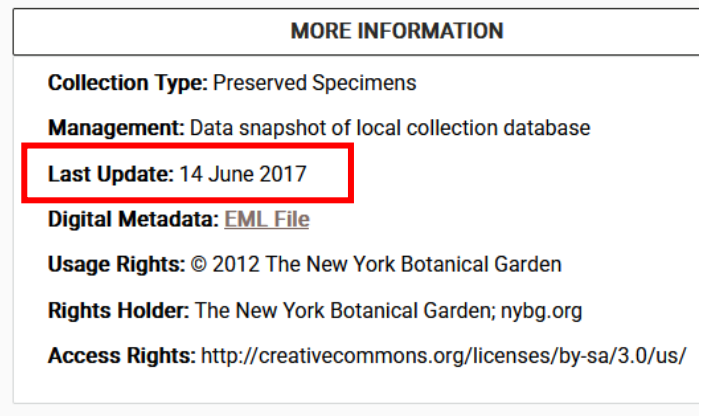
Feedback Loop:

- **CAUTION:** Annotations made to records via aggregators may be delayed in being received and resolved by the data provider, especially when made as comments in any of the many Symbiota Portals, unless the data are managed live in the portal.









Data Synchrony & Feedback Loop

If possible, leave comment at the record's source, or email the person(s) listed as the contact for the collection.

NYBG

Mobilization:
Integrated Publishing Toolkit
(IPT)

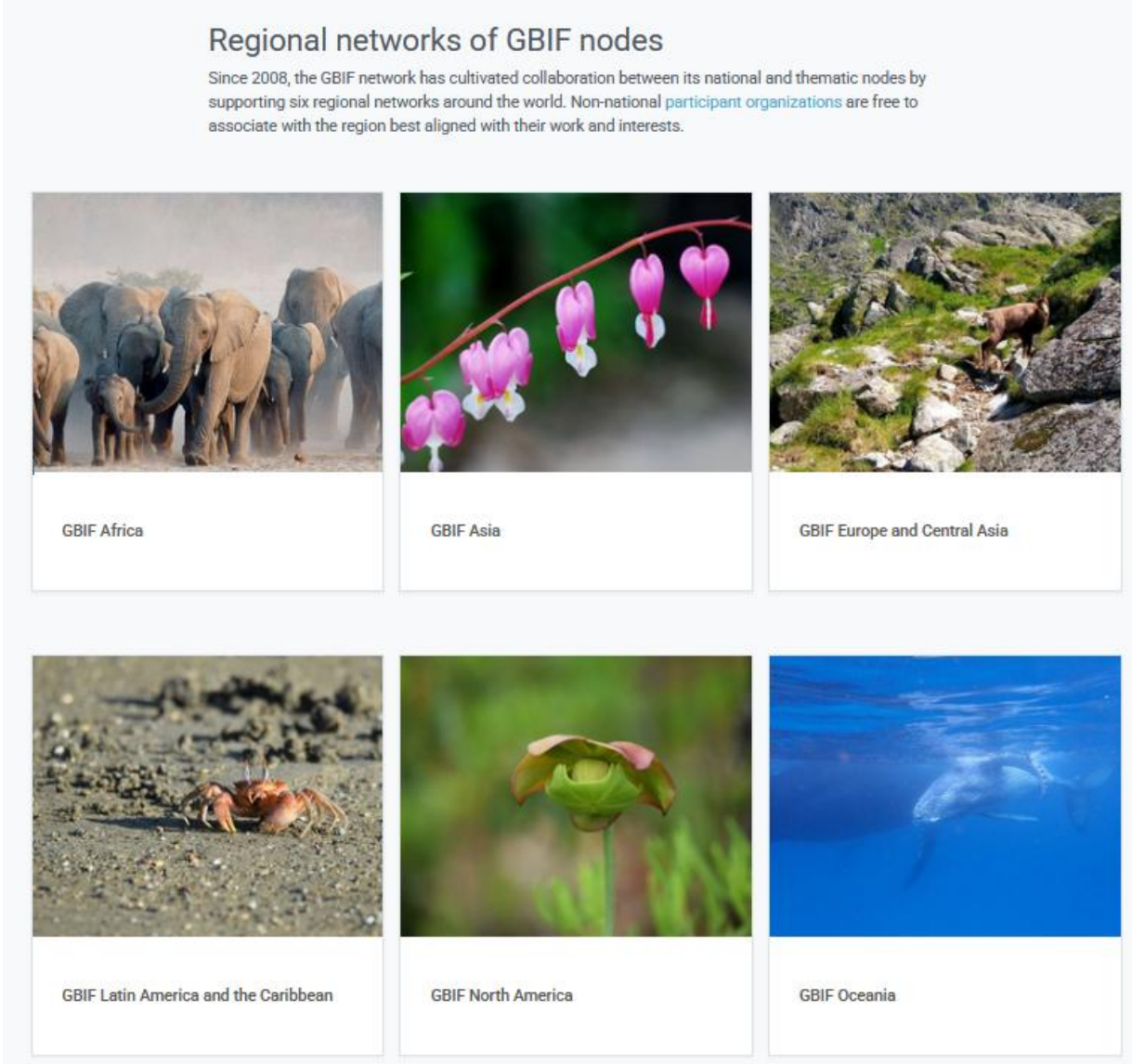
NYBG

GBIF Integrated Publishing Toolkit (IPT)

Free open-source software to create and manage repositories for sharing biodiversity datasets.

3 Modes of use:

- **Self-hosted**
- **National/thematic node installations:** coordinated GBIF-related activities for a country/community hosted for use by data publishers within their network
- **Regional cloud-hosted installations:** IPT instances with shared hard/software and storage services. Data-hosting centres



https://www.gbif.org/ipt

NYBG

Workflow

STAGE 01

Schedule Database Exports

Map fields in database to terms in Darwin Core Standard

Schedule database to run regular exports (e.g., weekly or monthly) in CSV format

Save CSV file on server

STAGE 02

Run a script to retrieve CSV and update MySQL database:

Set up a MySQL database a on server. (Alternatives: MS SQL server, Oracle, PostgreSQL, Sybase)

Fetch the exported CSV file from server

Update data in the MySQL database using any scripting language that runs on a scheduled basis

STAGE 03

Configure IPT

Download and install the GBIF IPT software on a server

Map MySQL database to the IPT resource by connecting IPT to your database using the appropriate database connection settings (e.g., JDBC)

Set up a user account within IPT for managing and publishing the dataset

STAGE 04

Set up Auto-Publishing Intervals in IPT

Configure automatic updates of the published dataset by setting a publishing interval (e.g., monthly, quarterly)

Test the auto-publishing functionality to ensure the data is correctly published to GBIF or other biodiversity portals

NYBG

Configure IPT

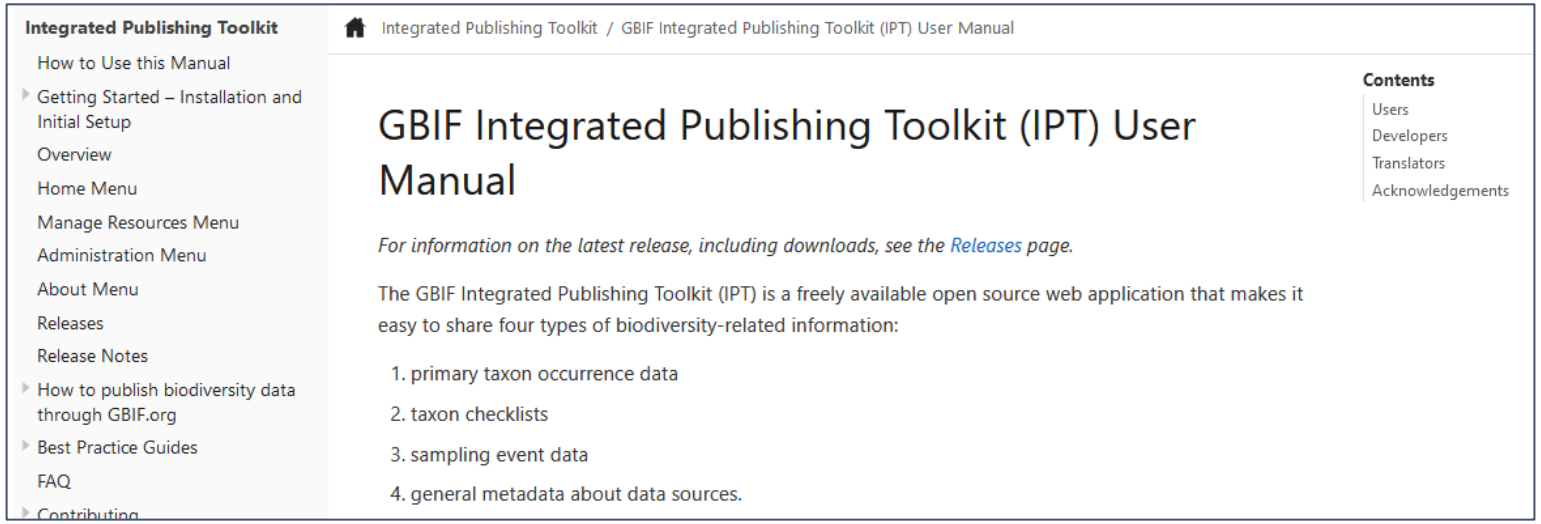
STAGE
03

Configure IPT

Download and install the GBIF IPT software on a server

Map MySQL database to the IPT resource by connecting IPT to your database using the appropriate database connection settings (e.g., JDBC)

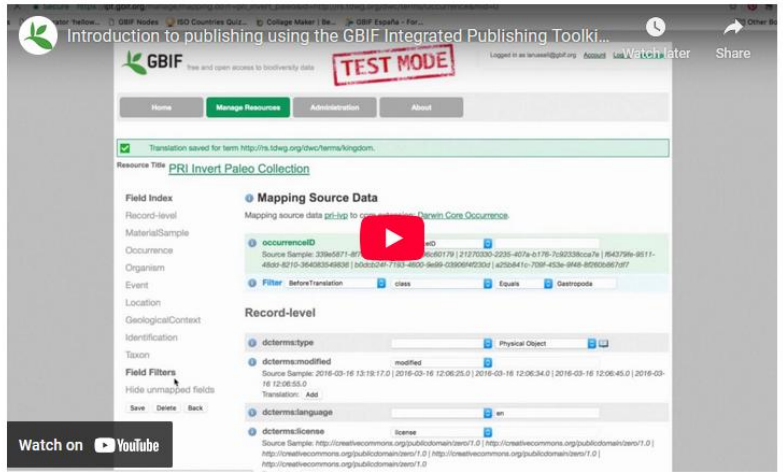
Set up a user account within IPT for managing and publishing the dataset



https://ipt.gbif.org/manual/en/ipt/latest/

Video [Intro to publishing using the GBIF IPT]

To understand how the IPT works, try watching this concise 25 minute live demo showing how a dataset gets properly published and registered through GBIF.org.



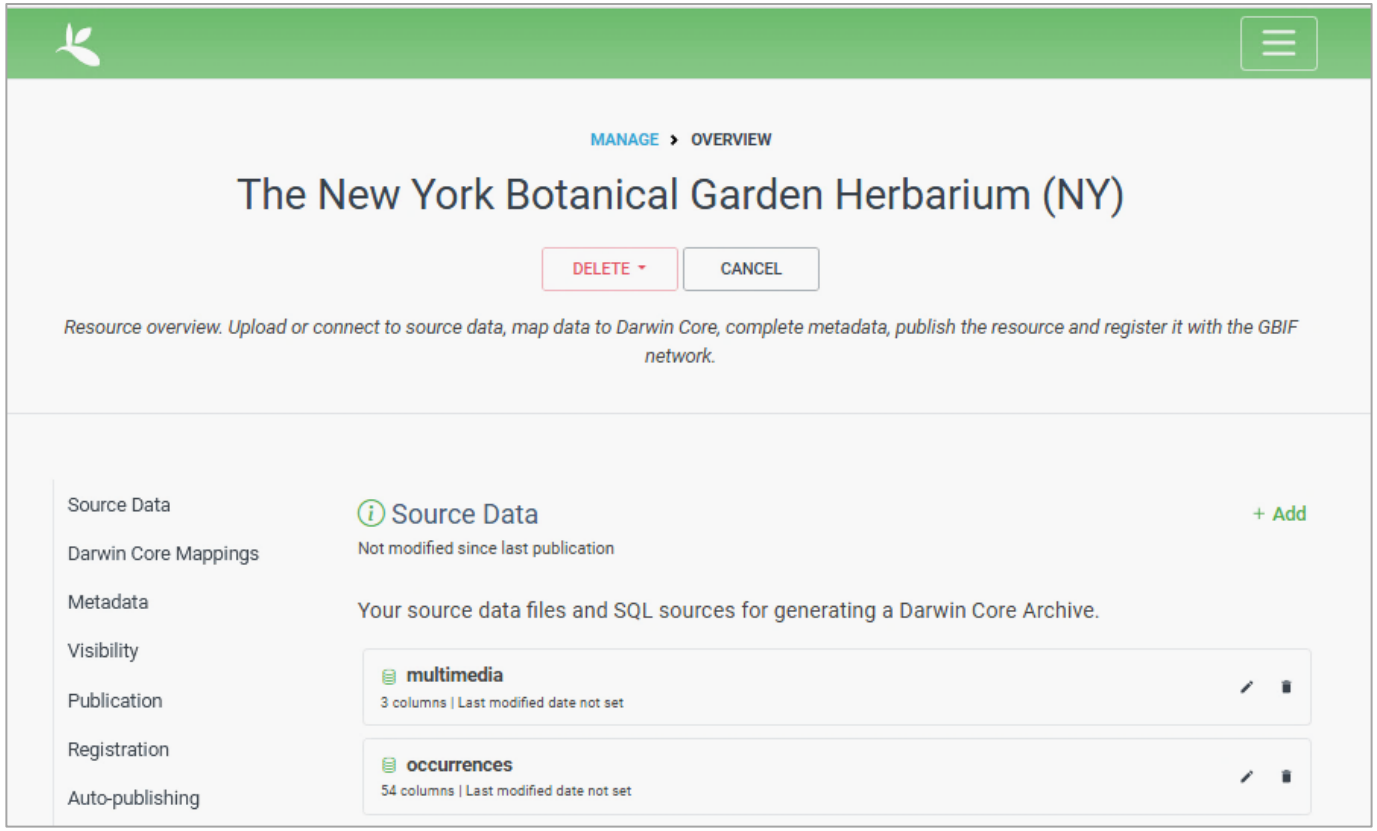
Auto-Publishing

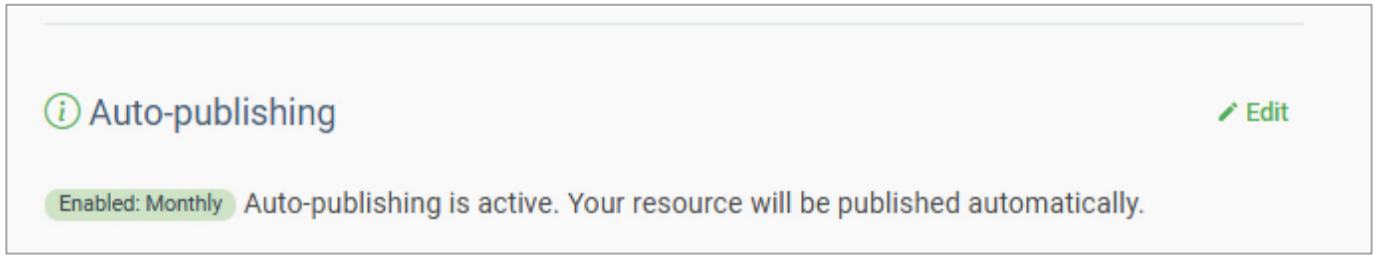
STAGE
04

Set up Auto-Publishing Intervals in IPT

Configure automatic updates of the published dataset by setting a publishing interval (e.g., monthly, quarterly)

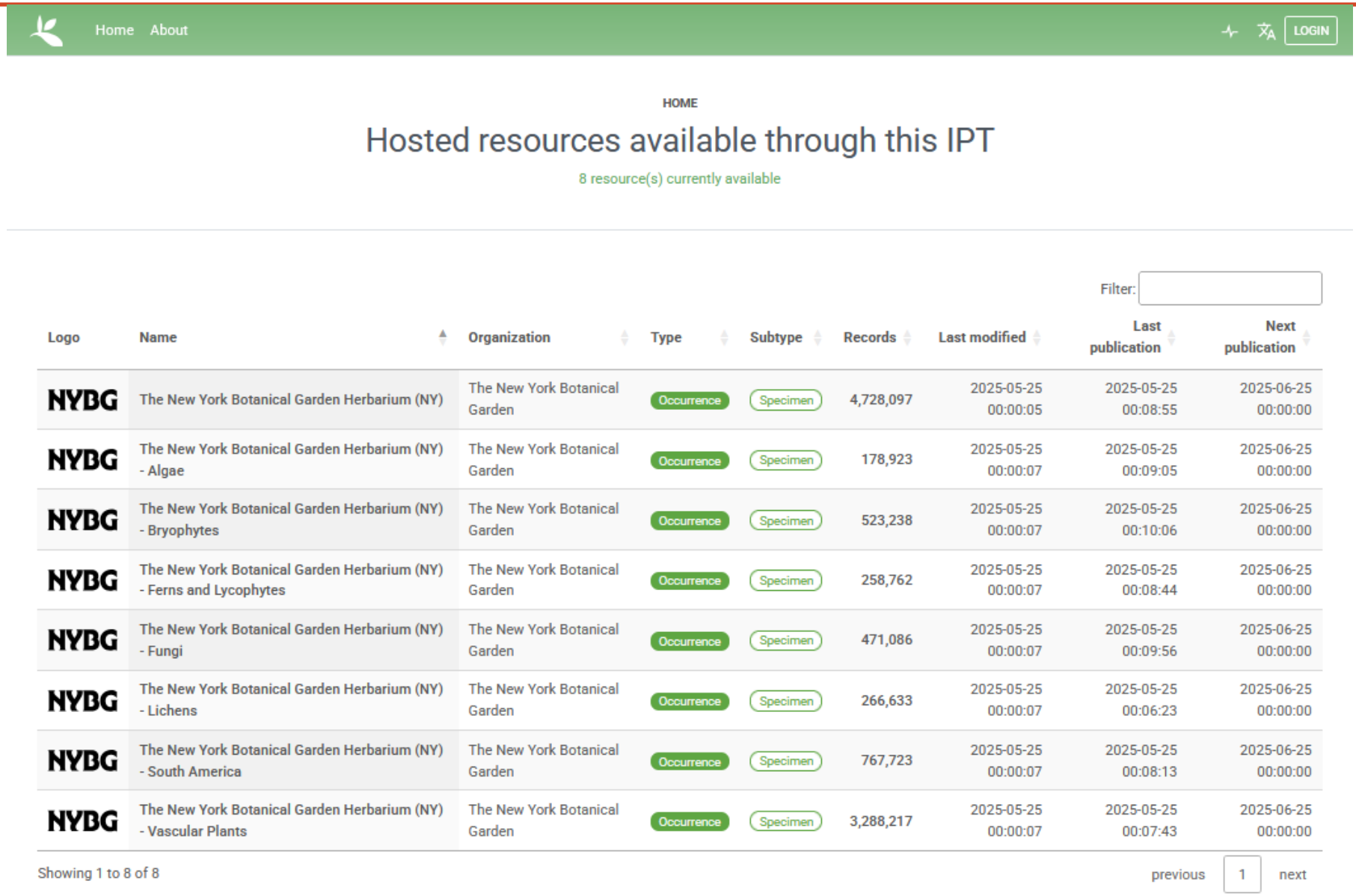
Test the auto-publishing functionality to ensure the data is correctly published to GBIF or other biodiversity portals





NYBG

Hosted Resources



NYBG

Mobilization:
Maximizing Data Impact with Minimal IT Support

NYBG

GitHub as a Simple Data Repository

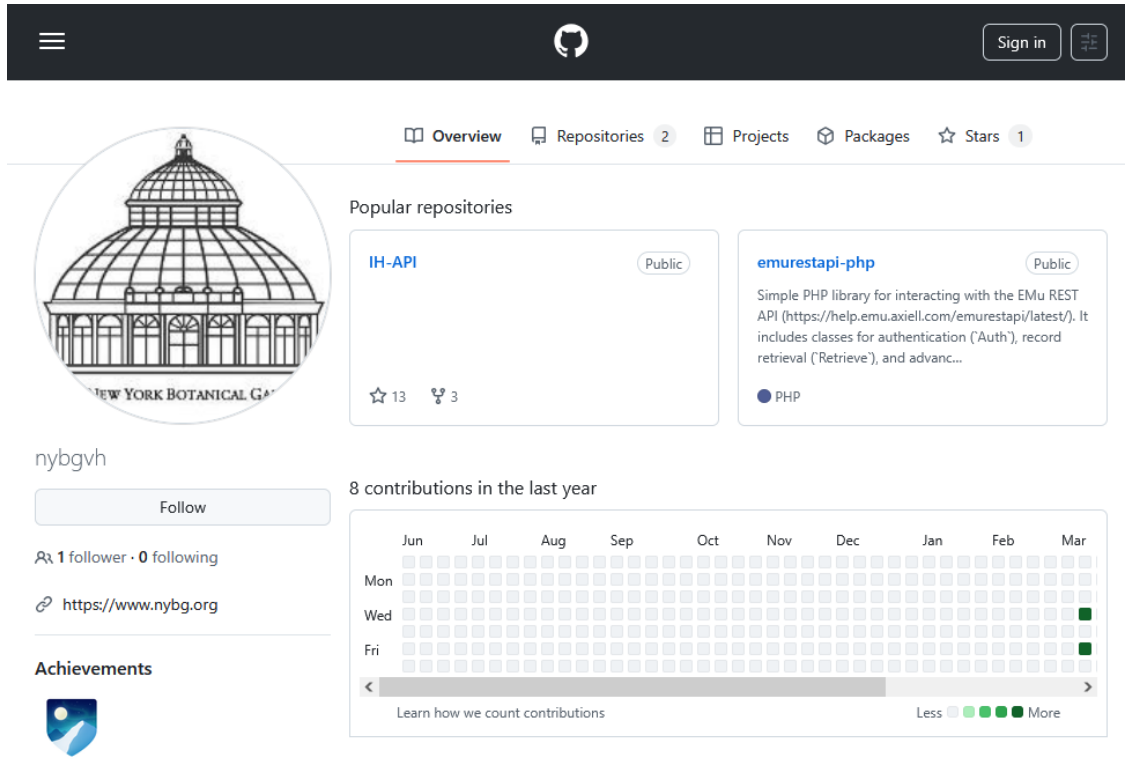
Free, globally accessible

No advanced technical setup required

Supports version control, coloration and transparency

How To:

- Create a free GitHub account: https://github.com/join
- Create a new repository (public or private)
- Upload your Darwin Core CSV Dataset
- Add a README.md file describing the dataset, license and usage
- Share the repository link with collaborators or the public



NYBG

GBIF Regional Cloud-Hosted IPT

GBIF's preferred platform for Darwin Core data publishing. No advanced technical setup required

Provides structured metadata, DOI assignment, and integration with GBIF.org

Managed by GBIF nodes to support institutions with limited IT capacity

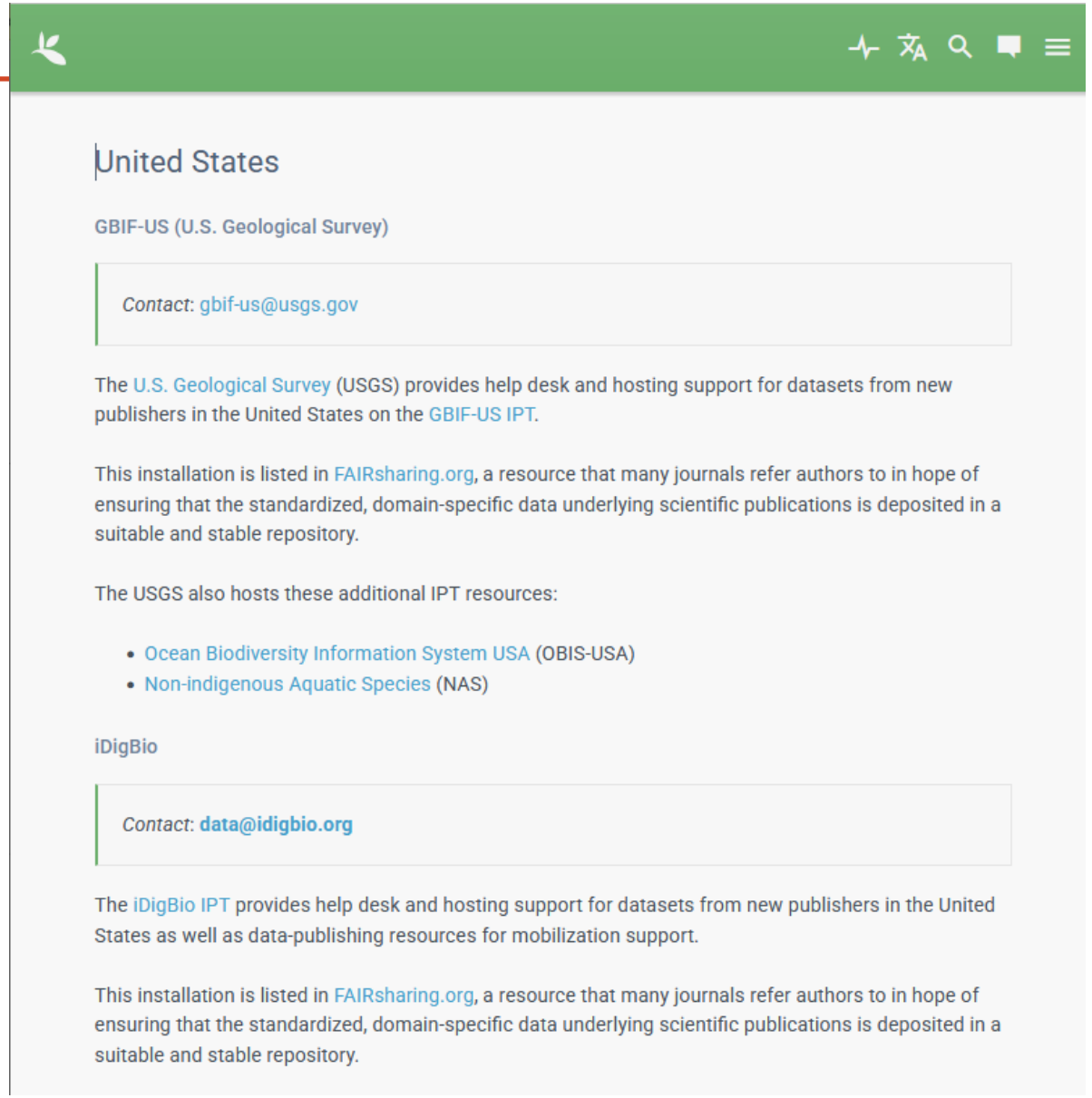
No need to host your own IPT instance

Learn more: https://www.gbif.org/ipt

Regional support avalable via GBIF nodes: https://www.gbif.org/the-gbif-network

- iDigBio IPT hosting (see Data Ingestion Guidance)

Email GBIF for hosting assistance: **helpdesk@gbif.org**



United States

GBIF-US (U.S. Geological Survey)

Contact: gbif-us@usgs.gov

The U.S. Geological Survey (USGS) provides help desk and hosting support for datasets from new publishers in the United States on the GBIF-US IPT.

This installation is listed in FAIRsharing.org, a resource that many journals refer authors to in hope of ensuring that the standardized, domain-specific data underlying scientific publications is deposited in a suitable and stable repository.

The USGS also hosts these additional IPT resources:

- Ocean Biodiversity Information System USA (OBIS-USA)
- Non-indigenous Aquatic Species (NAS)

iDigBio

Contact: **data@idigbio.org**

The iDigBio IPT provides help desk and hosting support for datasets from new publishers in the United States as well as data-publishing resources for mobilization support.

This installation is listed in FAIRsharing.org, a resource that many journals refer authors to in hope of ensuring that the standardized, domain-specific data underlying scientific publications is deposited in a suitable and stable repository.

NYBG