



How To Get Your Data To iDigBio

Joanna McCaffrey, iDigBio Biodiversity Informatics Manager
Summit 2016

Tuesday, 1 November 2016, Chattanooga, TN

[https://www.idigbio.org/wiki/index.php/
Data_Ingestion_Guidance](https://www.idigbio.org/wiki/index.php/Data_Ingestion_Guidance)



What do we mean by publishing data?

making biodiversity data publicly accessible & discoverable, in a standardized form, via a URL that is reproducible and automated.



The relevant verb: to publish


- No *sending*, No *pushing* – only *putting*
- No action occurs except when you first send me an email to say ‘my data are ready and you can find it here’.
- Once we know the ‘here’, we know to pick up any updates **AS LONG AS YOU RE-PUBLISH**

DATA Method #1


- What you already make available to GBIF
 - Using Darwin Core field names
 - Packaged in a Darwin Core Archive (DwC-A)
 - On an RSS feed (produced by **IPT***)
- ***[Integrated Publishing Toolkit]**



DATA Method #1B

- When you mark your data to publish, all the necessary parts of the package are generated.  *Symbiota* Promoting Bio-Collaboration
 - Custom Darwin Core Archive (DwC-A) on an RSS feed produced or updated by Symbiota
 - And almost automatic media
 - **<http://symbiota.org/docs/darwin-core-archive-data-publishing/>**

LepNet searchable datasets



Lepidoptera of North America Network








Home Search Images Fauna Projects Statistics Other Networks Work with LepNet Symbiota Contact Log In New Account Site


Home >> Collections

Specimens & Observations Specimens Observations

Select/Deselect All

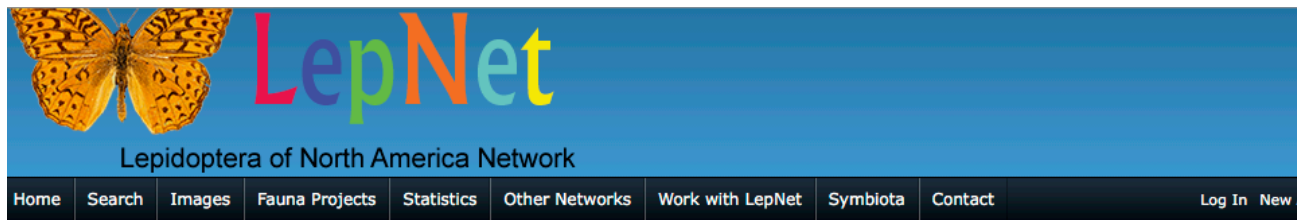
Live Data Collections

	<input checked="" type="checkbox"/> Academy of Natural Sciences Entomology Collection - Live Data (ANSP-ENT) more info
	<input checked="" type="checkbox"/> Arizona State University Hasbrouck Insect Collection (ASU-ASUHC) more info
	<input checked="" type="checkbox"/> BLM Mother Lode Field Office: The Bees of Pine Hill Preserve (BLM-MLFO) more info
	<input checked="" type="checkbox"/> Brigham Young University Arthropod Museum (BYU-BYUC) more info
	<input checked="" type="checkbox"/> C.P. Gillette Museum of Arthropod Diversity (CSU-CSUC) more info
	<input checked="" type="checkbox"/> Colorado Plateau Museum of Arthropod Biodiversity (NAUF-CPMAB) more info
	<input checked="" type="checkbox"/> Denver Botanic Gardens Collection of Arthropods (DBG-DBGA) more info
	<input checked="" type="checkbox"/> Denver Museum of Nature & Science (DMNS-DMNS) more info



Have you published?

- <http://symbiota4.acis.ufl.edu/scan/lepnet/portal/collections/datasets/datapublisher.php>



Home >> Sitemap >> **Darwin Core Archive Publisher**

Darwin Core Archive Publishing

The following downloads are occurrence data packages from collections that have chosen to publish their complete dataset as a Darwin Core Archive (DwC-A). A DwC-A file is a single compressed ZIP file that contains one to several data files along with a meta.xml document that describes the content. The archives contain three comma separated (CSV) files containing occurrences, identifications (determinations), and image metadata. Fields within the occurrences.csv are defined by the Darwin Core exchange standard. The identification and image files follow the DwC extensions for those data types.

Data Usage Policy:

Use of these datasets requires agreement with the terms and conditions in our Data Usage Policy. Locality details for rare, threatened, or sensitive records have been redacted from these data files. One must contact the collections directly to obtain access to sensitive locality data.

RSS Feed: <http://symbiota4.acis.ufl.edu/scan/lepnet/portal/webservices/dwc/rss.xml>

No data archives have been published for this collection

DATA #2

- Export your data as CSV/TXT file with DwC fieldnames & let us host it on our IPT or VertNet's
 - This method requires someone to have the time and skill to create a DwC file on their own.
 - We will help you register with GBIF

3 ways to get media to iDigBio:

- Use Audubon Core extension in IPT
 - Linked to the specimen
- **Via Symbiota**
 - Linked to the specimen
- Media appliance
 - Can be linked to the specimen

Media Metadata

- id (coreid of the specimen it links to)
- identifier (its own GUID)
- format (image/jpeg)
- accessURI (public path to your best quality jpg)

Symbiota/IPT Users - dataset naming

- Give it a complete name, institution, collection/
herbarium
- Description of the collection – what is in THIS
data
- Good contacts - the person who will respond to
requests

Example of good naming (2)

University of Vermont, Pringle Herbarium, North American bryophytes

Specimen Records: 17,697

Media Records: 16,941

iDigBio Last Ingested Date: 2016-07-28



The Pringle Herbarium (VT) contains 300,000 specimens, including vascular plants, bryophytes, lichens, algae and fungi. This portal contains our North American bryophyte specimens, numbering about 18,000. Other digitization projects cover type specimens, vascular plant specimens, North American lichens, macroalgae and macrofungi. These images and data are available through various portals. The herbarium does not maintain its own online database.

Contacts

Name *none*

Role *none*

Email CNABHadmin@asu.edu

Name Dorothy Allard, Virtual Herbarium
Coordinator

Role *none*

Email djallard@uvm.edu

DATASET INFO: rights

- Use Creative Commons standards:

– CC0 for data (not copyrightable)



– CC BY for media (at least)



IDENTIFIERS

- Every specimen and media record needs an identifier.
- We like UUIDs with a prefix:
`urn:uuid:2d5d3a8f-7a18-4825-a129-4a32b4ae58b8`

DATASET INFO: info about the provider (metadata)

Include your dataset **metadata** with your provider information (eml.xml):

- responsible parties (name, address, email, role)
- institution name, institution code, logo
- URL to the data at your institution
- descriptive paragraph about the institution, collection, and the dataset

DATASET INFO: update collections list

- iDigBio Collections

<https://www.idigbio.org/portal/collections>

- Index Herbariorum (Botany)

<http://sweetgum.nybg.org/ih/>

- GRBio.org Repositories:

<http://grbio.org/find-biorepositories>

Do you know what your **institutionCode** is?



Join the Community

- Join the Symbiota working group – community, webinars

Data Quality: Consider searchability in the aggregate

Dates – dwc:eventDate, dwc:day, dwc:month, dwc:year:

- this is not a month: Spring
- this is not a day: 10-18
- this is not a year: 1989? Or [1989]

Taxonomy – fill in dwc:scientificName, parse out the elements, fill in higher taxonomy

- this is not a species: shrimp, daisy

Tics: * [] {} ?

- Use the verbatim and remarks fields for things that do not fit the definitions.

Data Quality: Grooming and tics

Your dataset **is no longer just for making labels**, there are other considerations for being digital, and out in the wild:

- 1) Put dates in ISO 8601 format, i.e., YYYY-MM-DD, e.g., 2015-09-17
- 2) Parse apart scientific name
- 3) Conversely, put the piece parts into a scientific name
- 4) Provide as much higher taxonomy as you feel comfortable with, fill in tribe, sub+super family, kingdom, division, class, order) get out of 'family' land.
- 5) Make sure lat and lon coordinates are in decimal, and no N, S, E, W
- 6) Do not export '0' "n/a" in fields to represent no value, e.g., lat or lon, height
- 7) put elevation in METERS units in the elevation field without the units (e.g., the fields dwc:minimumElevationInMeters and dwc:maximumElevationInMeters already assume the numeric values are in meters, so there no need to include the units with the data)
- 8) And not to get too esoteric, do not use un-escaped newline characters or embedded tabs
- 9) Watch out for diacritics, save in UTF-8

à á â ã ä å

When is my work done?

- Digitization is never done
 - Label data
 - Georeferenced
 - Image
- Not until *your data are in iDigBio.*
 - It is not enough to get to it to Symbiota
 - Publish, re-publish with updates



Get involved!



idigbio.org/wiki



facebook.com/iDigBio



twitter.com/iDigBio



vimeo.com/iDigBio



idigbio.org/rss-feed.xml



idigbio.org/events-calendar/export.ics

I Dig Bio
do you?



iDigBio
Integrated Digitized Biocollections

