How To Get Your Data To iDigBio

Joanna McCaffrey, iDigBio Biodiversity Informatics Manager
Summit 2016
Tuesday, 1 November 2016, Chattanooga, TN
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.
Why publish data? The 4 biggies of data aggregation

ACCESSIBILITY
Data Use
Data Quality
Attribution
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.
  - Custom Darwin Core Archive (DwC-A) on an RSS feed produced or updated by Symbiota
  - And almost automatic media
  - READ: http://symbiota.org/docs/darwin-core-archive-data-publishing/
LepNet searchable datasets

Lepidoptera of North America Network

Live Data Collections

- Academy of Natural Sciences Entomology Collection - Live Data (ANSP-ENT)
- Arizona State University Hasbrouck Insect Collection (ASU-ASUHIC)
- BLM Mother Lode Field Office: The Bees of Pine Hill Preserve (BLM-MLFO)
- Brigham Young University Arthropod Museum (BYU-BYUC)
- C.P. Gillette Museum of Arthropod Diversity (CSU-CSUC)
- Colorado Plateau Museum of Arthropod Biodiversity (NAUF-CPMAB)
- Denver Botanic Gardens Collection of Arthropods (DBG-DBGA)
- Denver Museum of Nature & Science (DMNS-DMNS)
Have you published?

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

<table>
<thead>
<tr>
<th>Name</th>
<th>none</th>
</tr>
</thead>
<tbody>
<tr>
<td>Role</td>
<td>none</td>
</tr>
<tr>
<td>Email</td>
<td><a href="mailto:CNABHadmin@asu.edu">CNABHadmin@asu.edu</a></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Name</th>
<th>Dorothy Allard, Virtual Herbarium Coordinator</th>
</tr>
</thead>
<tbody>
<tr>
<td>Role</td>
<td>none</td>
</tr>
<tr>
<td>Email</td>
<td><a href="mailto:djallard@uvm.edu">djallard@uvm.edu</a></td>
</tr>
</tbody>
</table>
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 your 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 ā ä â̈ 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