Getting Your Data Published: Sending Data to iDigBio

Joanna McCaffrey, iDigBio Biodiversity Informatics Manager

TORCH Workshop
Botanical Research Institute of Texas (BRIT)
Saturday, 13 August 2016, Fort Worth TX

https://www.idigbio.org/wiki/index.php/Data_Ingestion_Guidance

iDigBio is funded by a grant from the National Science Foundation's Advancing Digitization of Biodiversity Collections Program (Cooperative Agreement EF-1115210). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

What do we mean by data publishing?

making biodiversity data publicly accessible & discoverable, in a standardized form, via a URL.

Why publish data? The 4 biggies for data aggregation

ACCESSIBILITY

Data Use

Data Quality

Attribution

Data publishing: where to begin with iDigBio?

- Email data@idigbio.org

- There are three ways to share data:

Least Ideal

Most Ideal

Technical skill vs. time, updatability, data buy-back etc.

DATA Method #1 – BEST

- What you already send to GBIF
  - Using Darwin Core field names
  - Packaged in a Darwin Core Archive (DwC-A)
  - On an RSS feed (produced by IPT)

iDigBio
Integrated Digitized Biocollections

DATA Method #2 – BETTER+

- 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 by Symbiota
  - almost automatic media
  - http://symbiota.org

Symbiota Promoting Bio-Collaboration

DATA #3 – GOOD ENOUGH

- Export your data as CSV/TXT file with DwC fieldnames & let us host it on our IPT or VertNet's

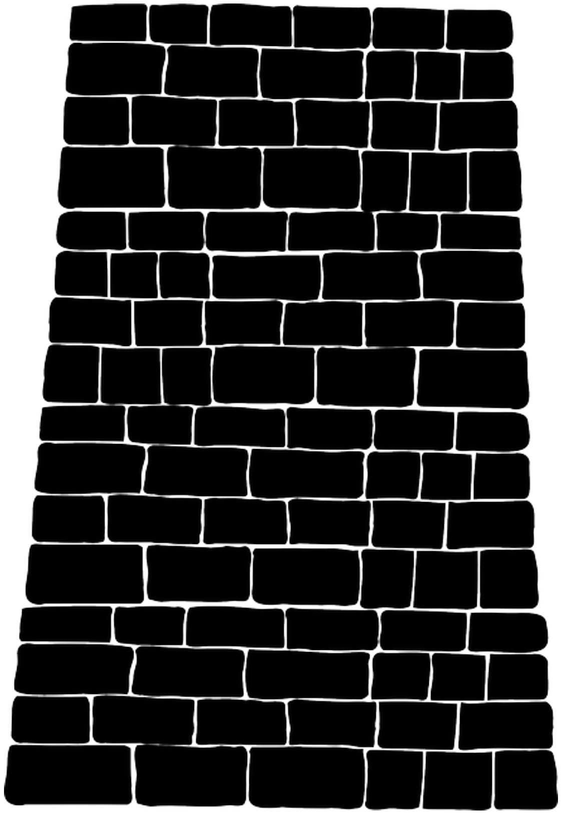
DATA #4 – Sub Par

- Throw the data over the wall and let us prepare it.
- Has its challenges:
  - data manipulations
  - UUID, data cleansing

⬇buy-back

⬇updates

⬇backlog



3 ways to get media to iDigBio:

1. Use Audubon Core extension in IPT	➢Linked to the specimen
2. Via Symbiota	➢Linked to the specimen
3. Media appliance	➢Can be linked to the specimen

DATASET INFO: info about the provider (metadata)

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

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

DATASET INFO: rights

Include data rights and rightsHolder information:

- Use Creative Commons standards:
  - – CC0 for data (not copyrightable)
  - – CC BY for media (at least)

iDigBio
Integrated Digitized Biocollections

DATASET INFO: update collections lists

- iDigBio Collections

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

- Index Herbariorum

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

- GRBio.org Repositories:

http://grbio.org/find-biorepositories

Do you know what your institutionCode is?

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' 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

Symbiota Notes

- Your dataset
    - 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
- Join the Symbiota working group – community, webinars

Data Management Plan

- Build a robust DMP – look at DataOne
  - Who will be contributing data (roles and responsibilities)?
    - With what software are they managing their data?
    - Metadata used (Darwin Core?)
  - How will they be mobilizing it
    - Dataset names
    - GUIDs (occurrenceID)
  - Record counts, media counts
  - How they will get it to iDigBio
  - Archiving strategy, backup protocol
  - Responsible parties
  - Other repositories (GBIF, VertNet)
  - Data extensions (e.g., associations)

Don’t Wait – Catch Up!

- Hundreds of herbaria (390+) are ahead of you in their digitization efforts
- Sharing data keeps you relevant
- TORCH: thank you so far !
  - TEX/LL (Tom & George)
  - ASU (Marcy) coming soon
  - BRIT (Jason - Bryophytes & Lichens)
  - CSU – (Clark - Macrofungi)

iDigBio
Integrated Digitized Biocollections

Thank you for your attention

www.idigbio.org

facebook.com/iDigBio

twitter.com/iDigBio

vimeo.com/idigbio

idigbio.org/rss-feed.xml

webcal://www.idigbio.org/events-calendar/export.ics

iDigBio is funded by a grant from the National Science Foundation's Advancing Digitization of Biodiversity Collections Program (Cooperative Agreement EF-1115210). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.