How To Submit Data to iDigBio for ‘TORCH VIII + iDigBio’

Joanna McCaffrey

May 2014, Alpine, (3.0)
Where to begin? It starts with a conversation

DATA #1 - BEST

What you already send to GBIF
using Darwin Core fieldnames
packaged in a Darwin Core Archive (DwC-A)
on an RSS feed produced by IPT
DATA #2 – BETTER+

Custom Darwin Core Archive (DwC-A) on an RSS feed produced by Symbiota

↑ automatic images
↓ narrower schema
DATA #3 – GOOD ENOUGH

A custom CSV or TXT file, with XML style field names from Darwin Core, e.g., domain:fieldName

dwc:catalogNumber
ac:provider

↓ personnel maintenance costs
DATA #4 - ADEQUATE

The last, and least preferable way: Throw the data over the wall and let us prepare it.

👇 buy-back
👇 updates
DATASET INFO : info about the provider

Send your dataset info with your provider information:

- responsible parties (name, address, email, role)
- institution name, institution code
- URL to the data at your institution
- descriptive paragraph of the collection
DATASET INFO : copyrights

Include data rights information
Use Creative Commons standards:
  CC0 for data (not copyrightable)
  CC BY for media (at least)
DATASET INFO : updating GRBio.org

GRBio.org
    Repositories:
    http://grbio.org/find-biorepositories
    Institutional collections:
    http://grbio.org/find-institutional-collections
IMAGES #1 – use Audubon Core extension to IPT

Create a file of Audubon Core metadata includes URL to images and camera info (EXIF), photographer, PLUS a link to the specimen record via occurrenceID

↑ hooked up to specimen
IMAGES #2 – via Symbiota

↑ hooked up to specimen
IMAGES #3

Image ingestion appliance

↓ not yet hooked up to specimen
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

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., 2014-06-22
2) Parse out 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
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
9) Watch out for diacritics, save in UTF-8
The End

Questions? Comments?