Data Standards & Mobilization

Joanna McCaffrey, iDigBio Biodiversity Informatics Manager

EMu NHSIG
University of Pennsylvania Museum of Archaeology and Anthropology
Wednesday morning, 7th October 2015, Philadelphia PA

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.

Do I need to mention data standards?

Having lived in EMu world for as long as you have, you are familiar with thinking beyond the label

- A robust data schema
    - Reserved vocabularies, a generous helping of rigour
- Mapping to the shadow DwC table in the catalog
- Sharing within your EMu world, but to share in the aggretate, you need something else.

iDigBio chose Darwin Core (DwC) to be our first choice for representing specimen data

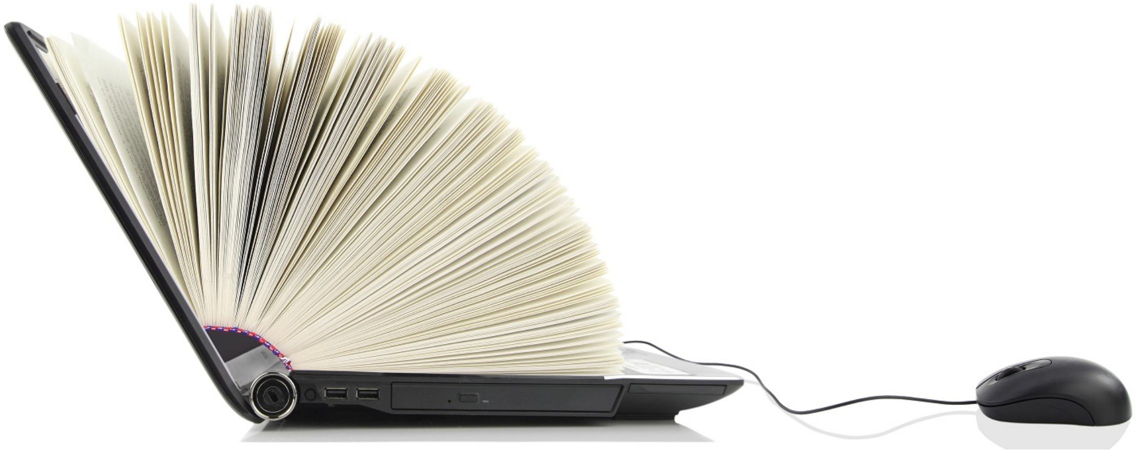
iDigBio
Integrated Digitized Biocollections

One other essential for data sharing

- A GUID - Globally Unique Identifier
  - Persistent
  - Robust
  - Example:
    - urn:uuid:f47ac10b-58cc-4372-a567-0e02b2c3d479
  - See our Data Ingestion Guidance document for details:
    - https://www.idigbio.org/wiki/index.php/Data_Ingestion_Guidance

What do we mean by mobilizing data?

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



The 4 biggies for data aggregation

ACCESSIBILITY

Data Use

Data Quality

Attribution

Copyright 2015

tagxedo.com

Data publishing: where to begin with iDigBio?

- Email data@idigbio.org

- There are several ways to share data:

Least Ideal

Most Ideal

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

Register your data

3 places to make sure your collections information is correct:

- 1) our collections list
  - https://www.idigbio.org/portal/collections
- 2) grbio.org
- 3) Index Herbariorum (NY) (if you are a herbarium)

Do you know what your institutionCode is?

Prepare: DATA Method #1 – BEST

- What you already send to GBIF
  - Using Darwin Core field names
  - Packaged in a Darwin Core Archive (DwC-A)
    - Darwin Core for specimen data
    - Audubon Core extension for media (.jpg, .stl)
  - On an RSS feed (produced by IPT)

Prepare: DATA Method #2 – BETTER+

- Any of you who are using Symbiota for a TCN:
  - 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
  - automatic media
  - http://symbiota.org

Prepare: DATA #3 – GOOD ENOUGH

- Export your data as CSV/TXT file with DwC fieldnames & let us host it on our IPT or VertNet's
  - e.g., domain:fieldName
    - dwc:catalogNumber
    - ac:provider

⬇ personnel maintenance costs

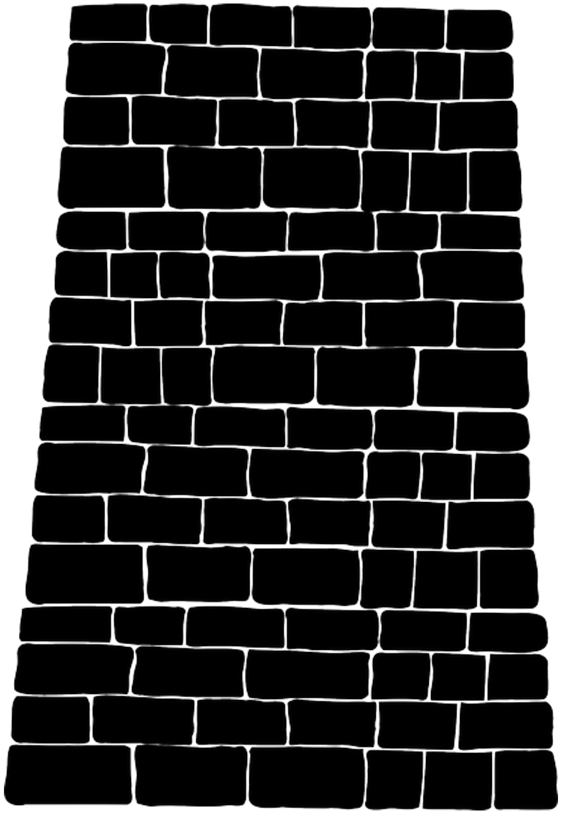
DATA #4 - ADEQUATE

- 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 ingestion appliance

   ➢Can be linked to the specimen

DATASET INFO: info about the provider

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 collection

DATASET INFO: rights

Include data rights and rightsHolder information:

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

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., 2015-09-17
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 or embedded tabs
9) Watch out for diacritics, save in UTF-8

à á â ã ä å

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.