Achieving a unified collection management system, preparing data

Joanna McCaffrey, iDigBio Biodiversity Informatics Manager
Scripps Institute of Oceanography
Wednesday, 3rd February 2016, La Jolla CA

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.

iDigBio
Integrated Digitized Biocollections

Why have a unified database ?

- Economies of scale, effort
- Sharing ideas, costs (IT staff)
- Going in a new direction together can invigorate staff
- Possibly introduce new technology along the way, e.g., data export

How do we get there?

- What follows are some strategies for achieving a unified database
  - Process
  - A journey together

Considerations for selecting a collections management system - motivation

Establish institutional motivation to unify databases

- earnest desire could be generated by a focus group with institutional stakeholders, funding generators, users of data, people who input data, data system supporters (curators), IT,

Considerations for selecting a collections management system – agreement 1

Document and agree on a priority feature set that is necessary versus desired:

a. system is extensible, customizable,
b. responsive vendor,
c. supports reports, auditing,
d. generates labels,
e. supports loans (partial returns, cataloged and uncataloged specimens),
f. supports pest management,
g. supports multimedia attachments (PDF loan forms, image, sound files, etc.),
h. supports web access and privacy,

Considerations for selecting a collections management system – agreement 2

i. all the input/output scenarios you might envision:
   - Import/export abilities, can it support Darwin Core field mappings
   - plan B scenario if software or the internal project becomes unfunded,

j. affordable user license costs: per seat, pool,

k. has basic, and easily customizable help,

l. Mac versus PC, perhaps an issue in your user population,

m. has a robust security model (passwords, users, groups, permissions, input and query defaults, controlled vocabularies),

n. supports accessibility, different character sets,

Considerations for selecting a collections management system - IT

- Proprietary, open source, hybrid, cloud-based
  - Who decides what features to develop?
  - Who does maintenance?

Considerations for selecting a collections management system

- Interest in having what your peers have: economies of training, user community,
- Beware of demo-ware

iDigBio
Integrated Digitized Biocollections

Considerations for selecting a collections management system

- Shop vendors and score them on their ability to meet necessary features above, with extra points for desired ones

Considerations for selecting a collections management system

- Get a full demo copy and enter data with a realistic test case dataset, score on ease of learning the system

Considerations for selecting a collections management system - costs

When choosing preferred system, consider costs derived from these sources:

1. upfront software costs,
2. software maintenance,
3. long term costs (server space, server replacement, backup),
4. where it is hosted,
5. IT support of system without being the bottleneck,
6. hidden costs of conversion, cleansing, improvements,
7. institutional bioinformatics staff support to continue development of data, ('data curator', biodiversity informatics manager).

https://www.idigbio.org/content/biological-collections-databases

Some software applications options

- These run the gamut from almost free -> pricey with annual licenses, and from little or no support -> lots of support. They all come with a large community of users.
- Symbiota
- Specify
- Arctos
- EMu

Preparing Data

- Discuss Creative Commons rights
  - CC0 for data (not copyrightable)

  - CC BY for media (at least)

You will need to define AC rights and rightHolder

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

Data cleaning tools

- Open Refine
- Remember Data Carpentry?
- Power user Excel
- Access VB scripts

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.