Getting Your Data Out There: Data Publishing & Data Standards with iDigBio

Molly Phillips & Joanna McCaffrey
data@idigbio.org
What do we mean by data publishing?
Why publish data?

Data Use

Data Quality

Attribution
Why Publish?
Data Standards
Biodiversity data standards

- **Darwin Core**
  biodiversity informatics (specimen and observation data)

- **Audubon Core**
  multimedia related to specimens
**Darwin Core**

**What:** Darwin Core is a glossary of terms intended to facilitate the sharing of information about biological diversity.

**How:** The Darwin Core is based on taxa, their occurrence in nature as documented by observations, specimens, samples, and related information.

**Where:** [http://rs.tdwg.org/dwc/terms/](http://rs.tdwg.org/dwc/terms/) provides reference definitions, examples, and commentaries.
Data standards & Darwin Core

• With data standards like Darwin Core, we have established rules for how we enter certain fields.

• examples:
  – Date
  – Lat/Lon
  – Genus
  – Species
Data Sharing
Data publishing: where to begin with iDigBio?

• Email data@idigbio.org

• There are four basic ways to share:

<table>
<thead>
<tr>
<th>Least Ideal</th>
<th>Most Ideal</th>
</tr>
</thead>
</table>

Technical skill vs. time, updatability, data buy-back etc.
# 1 – BEST:
Send data to GBIF Great, we’ll take that!

- Darwin Core Archive (DwC-A)
- on an RSS feed produced by IPT
- [https://code.google.com/p/gbif-providertoolkit/](https://code.google.com/p/gbif-providertoolkit/)
#2- Also great: Use Symbiota

• 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](http://symbiota.org)
# 3- Adequate: 
Export your data as CSV/TXT file with DwC fieldnames & let us host it on our IPT

- Create a custom CSV or TXT file,
  - with XML style field names from Darwin Core,
    - e.g., domain:fieldName
    - dwc:catalogNumber
    - ac:provider
# 4- Will work in a pinch: Throw your data over the wall

• This method has its challenges:
  – data manipulations
    • UUID, higher taxa, dates, zeros...
  – Updates
  – Buy-backs
  – Backlog
Media

3 ways to get media to iDigBio:

1. use Audubon Core extension to IPT ➢ Linked to the specimen

2. via Symbiota ➢ Linked to the specimen

3. Media ingestion appliance ➢ Can be linked to the specimen
Metadata
Metadata

A set of data that describes and gives information about other data.

– For us, its data that describe a biodiversity dataset.

• Metadata facilitates:
  – Data discovery
  – Search & retrieval
  – Reuse (licensing)
  – Attribution
  – Expressions of fitness-for-use
  – Communication
What metadata does iDigBio need?

• Information about the provider
  – responsible parties (name, address, email, role)
  – institution name, institution code
  – URL to the data at your institution
  – descriptive paragraph of the collection

Equivalent to the eml.xml file produced by IPT
Check for existing collections:

In GRBio.org

• Repositories:
• http://grbio.org/find-biorepositories

• Institutional collections: http://grbio.org/find-institutional-collections
Copyrights: please include rights info

Use:

- CC0 for data (not copyrightable)
- CC BY for media
Data Ingestion
What happens when you send us your data?

iDigBio Data Flow Diagram

Collections
Specify, EMu, Symbiota, ...

Publishers
IPT, Symbiota, iDigBio Feeder

Data Ingestion
Python, PostgreSQL, JSON, Redis

iDigBio API
PostgreSQL, Riak

Searchable Indexed Data
Elasticsearch

iDigBio Portal Web Site
HTML5, jQuery, Backbone, Node.js, Express
https://www.idigbio.org/portal

Scientific Community
(Researchers, Scientists, Developers, and downstream consumers)
Further Resources…

• [https://www.idigbio.org/portal/publishers](https://www.idigbio.org/portal/publishers) look who is already providing data to iDigBio.
• [http://rs.tdwg.org/dwc/terms/](http://rs.tdwg.org/dwc/terms/) for the Darwin Core glossary.
• [https://www.idigbio.org/wiki/images/0/01/ImageIngestionCheatSheet_Sheet1.pdf](https://www.idigbio.org/wiki/images/0/01/ImageIngestionCheatSheet_Sheet1.pdf) tips on using iDigBio’s image ingestion appliance.
• [https://www.idigbio.org/wiki/images/0/03/GUIDgeneration.pdf](https://www.idigbio.org/wiki/images/0/03/GUIDgeneration.pdf) how to create UUID GUID in an excel spreadsheet.
• [https://www.idigbio.org/wiki/images/e/e2/ToPrepareAnAudubonCore.pdf](https://www.idigbio.org/wiki/images/e/e2/ToPrepareAnAudubonCore.pdf) how to prepare an Audubon Core file using IPT.
• [https://code.google.com/p/gbif-provider-toolkit/](https://code.google.com/p/gbif-provider-toolkit/) more information about the GBIF IPT.
• [http://symbiota.org](http://symbiota.org) Symbiota.
Thank you!

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.
Reserved fields & Darwin Core

• **Dates:** – `dwc:eventDate` is a date and nothing else
  – Also for `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 are reserved fields too:**
  – this is not a species: shrimp

Use the verbatim & remarks fields for things that do not fit the definitions.
More data tips…

1) Put dates in ISO 8601 format, i.e., YYYY-MM-DD, e.g., 2014-06-22
2) Fill in dwc:scientificName with genus and species
3) Parse out the dwc:scientificName elements to fill in dwc:genus and dwc:specificEpithet
4) Provide as much higher taxonomy as you 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 not degs, mins, secs,
6) Do not export '0' in fields to represent no value
7) Get rid of your tics: * [] {} ?...
8) Put elevation in METERS units in the elevation field without the units. Watch out for diacritics, save in UTF-8 (encoding)
Data ingestion process

- Ingesting:
  - Ingested, update every Wednesday
  - Note: once submitted, cannot be withdrawn

- Evaluating:
  - Error with ingestion, evaluate
  - Fix internally, re-submit

- Mobilizing:
  - Reply: Submit data for preparation/inspection
  - Return data to provider to fix

- Negotiating:
  - Pre-negotiating: email, phone call, conversation
Architecture components

iDigBio Specimen Portal
HTML5, jQuery, React, Node.js, Express

iDigBio Search API
Elasticsearch

Text Indexes
Elasticsearch

SQL Indexes
PostgreSQL

Bulk Text Storage
Riak

Appliances
KVM, Xen, VirtualBox, VMware

iDigBio Metadata API
Node.js, Restify, REST, JSON

iDigBio Object API
Python, CherryPy, REST, JSON

Third Party API Consumers
Python, jQuery

iDigBio Worker Services
Apache, CherryPy, Python, Celery

Cloud Node
Cloud Node
Cloud Node
Cloud Node
Cloud Node
Cloud Node

Cloud Node
Cloud Node
Data collection & standards

- Data quality starts with what you collect & ends with what you publish