

# **Preparing Data For Export**

Darwin Core Archive, IPT, GUIDs



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.



## **Preparing Data for Export**

- Data in standard format, using accepted metadata, parsed out as far as you can go
  - Darwin Core, Audubon Core
- Set up a packaging mechanism once, reproducible
  - IPT
- Share widely
  - iDigBio, VertNet
- Register for optimal reach
  - GBIF



## Why bother?

- Quality is #1
  - With more eyes on your data, it gets better
- Grants
  - Funders want to know that you are sharing your data – digitization and mobilization
- Another way to get credit for your work
- Part of modernizing your collection practice
  - People will want to work with you, for you



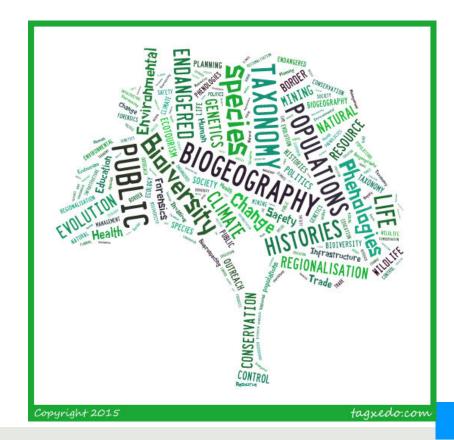
## The 4 biggies for data aggregation

ACCESSIBILITY

Data Use

Data Quality

Attribution





### GBIF has IPT and instructions for installation

- Manual: https://code.google.com/archive/ p/gbif-providertoolkit/wikis/ IPT2ManualNotes.wiki
- Vimeo: <a href="https://vimeo.com/116142276">https://vimeo.com/116142276</a>
- Other experts in the community
- takes 30 minutes, and occasional updates



## IPT – Integrated Publishing Toolkit

- Your go-to tool for constructing a Darwin Core Archive (DWC-A)
  - a ZIP file containing
    - Occurrence records\*
    - Multimedia records (in Audubon Core)
    - Identifications
    - EML\*
    - Meta\*
  - \* required



Name	↑ Date Modified	Size	Kind
eml.xml	1/26/16	8 KB	eXtensibleML) document
meta.xml	1/26/16	4 KB	eXtensibleML) document
multimedia.txt	1/26/16	6 MB	Plain Text Document
occurrence.txt	1/26/16	21.2 MB	Plain Text Document

#### **Contents of Darwin Core Archive with media**

2 text files and 2 XML files



### **IPT - 2**

- Can be set up to do a dynamic query of your database, or accept a CSV export
- Interactive tool to map your fields to Darwin Core terms
- Possible sources of IPT support, if you are not putting up one yourselves:
  - VertNet
  - iDigBio



Home Hosted resources available through this IPT Filter: Name Logo Organisation Type Subtype Records Last Last Next publication publication modified AMNH Invertebrate Paleontology AMNH Occurrence Specimen 976 2015-12-2015-12-10 **\$** 10000000 10 Collection **BUT Friesner Herbarium** BUTLER BUT Occurrence Specimen 42,745 2016-01-2016-01-26 2016-02-02 27 15:15:40 SDSM Invertebrate Paleontology Not registered Occurrence Specimen 1,134 2015-12-2015-10-10 Collection 22 USF USF Herbarium Not registered Occurrence Specimen 117,032 2016-01-2016-01-26 2016-02-02 26 09:30:39 UTEX Culture Collection of Algae at The UTEX Occurrence Specimen 2,813 2016-01-2016-01-13 University of Texas Living Algae 13 Holdings Showing 1 to 5 of 5 previous next The most recently updated resources are also available as an RSS feed. .

#### iDigBio IPT example

http://ipt.idigbio.org



### **IPT - 3**

- Includes
  - RSS feed (Really Simple Syndication) for updates
  - GBIF registration possibilities



## Globally Unique Identifiers (GUIDs)

- Single <u>most important</u> piece of info for each record that you export
- Identifiers that a born with the digitial record of your occurrence or multimedia records.
- Distinct from catalog number, accession number, barcode, etc.
- A way for providers to send data out to aggregators and always be able to identify a record as theirs.
- It is persistent



### **GUID - 2**

- A GUID is never
  - Deleted
  - Re-used

• Use same-as relationship if you change identifiers along the way



### **GUID - 2**

- Some formats:
  - Darwin Core triplet (old school) with prefix

<dwc:institutionCode>:<dwc:collectionCode>:<dwc:catalogNumber>

urn:catalog:SIO:Fish:122



### **GUID - 3**

- UUID (new school)
  - f47ac10b-58cc-4372-a567-0e02b2c3d479
  - urn:uuid:f47ac10b-58cc-4372a567-0e02b2c3d479
- ark:/87286/f47ac10b-58cc-4372a567-0e02b2c3d479



## Guidance on getting data to iDigBio

 https://www.idigbio.org/wiki/index.php/ Data\_Ingestion\_Guidance



# Thank you for your attention





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.