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
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

• Vimeo: https://vimeo.com/116142276
• 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
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
## 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 **most important** piece of info for each record that you export

• Identifiers that are born with the digital 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

  \texttt{<dwc:institutionCode>:<dwc:collectionCode>:<dwc:catalogNumber>}

  \texttt{urn:catalog:SIO:Fish:122}
GUID - 3

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