CYWG (CyberInfrastructure Working Group)
October 2014

iDigBio Data Ingestion

Dan Stoner
Advanced Computing and Information Systems Laboratory (ACIS)
University of Florida

✉ dstoner@acis.ufl.edu
@thatlinuxbox

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. All images used with permission or are free from copyright.

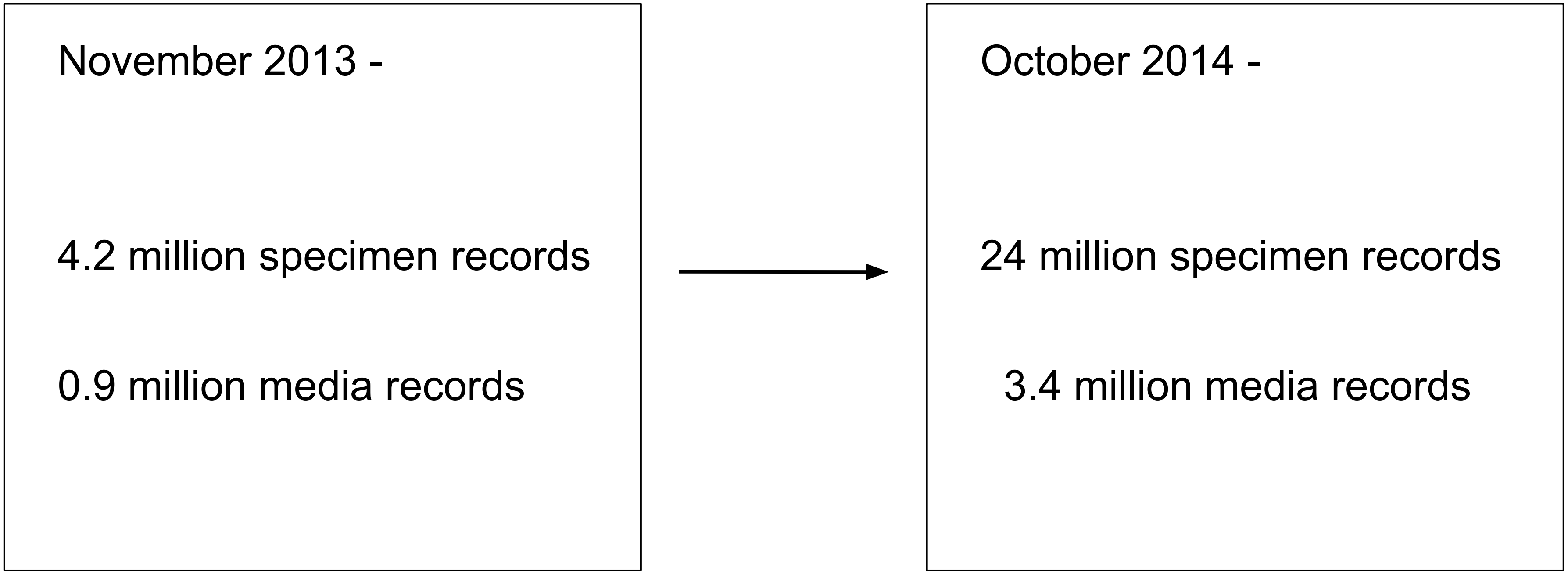
Over 300 Data Providers...

Sigillum Universitatis Arizonensis 1885
A
ASU HIC
BYU Brigham Young University
Boise State University
UCO
Dickinson State University
Duke University
Eastern Illinois University
Eastern New Mexico University
Evergreen The Evergreen State College Olympia, Washington
Fort Lewis College
The Field Museum
University of Hawai'i
Ada Hayden Herbarium
KU Biodiversity Institute The University of Kansas
Cincinnati Museum Center
Cornell University Founded A.D. 1865
The Ohio State University
Montana State University
M
MSC
Norton-Brown Herbarium
NCSM Fishes Unit
North Dakota State University
NYS M
OSU Oregon State University
Rocky Mountain Herbarium
Rutgers The State University of New Jersey
Sam Noble Museum
San Francisco State University 1899
SIU Carbondale
T
The University of Alabama Ichthyological Collection
UCSC Fungal Herbarium
M
UNLV
University of Richmond
USF University of South Florida
The Burke Museum of Natural History and Culture University of Washington
Virginia Polytechnic Institute and State University
V State

Black Hills State University
USDA
Plant to planet.
MUSEUM NATURAL HISTORY
CUP
D&E Davis & Elkins College
NATURE & SCIENCE
ENLC
HARVARD UNIVERSITY HERBARIA
FLORIDA MUSEUM OF NATURAL HISTORY
KU GPSC
BENEDICTINE COLLEGE
PLH
ILLINOIS NATURAL HISTORY SURVEY
LSU
MAINE
COLORADO MESA UNIVERSITY
MICHIGAN
WESTERN ILLINOIS UNIVERSITY
Colorado Plateau Museum of Arthropod Biodiversity
NCS LG
UNIVERSITY OF NEBRASKA STATE MUSEUM
THE UNIVERSITY OF NEW HAMPSHIRE 1866
NATURAL HISTORY MUSEUM OF UTAH
NM STATE UNIVERSITY
NY BG
Biology
The Academy of Natural Sciences PH Philadelphia
ROCKY MOUNTAIN HERBARIUM
TEAM SCANAR
South Dakota State University SDSU
ESF
UT
UCMP
UC
UDAF
THE UNIVERSITY of NEW MEXICO
UPPSALA UNIVERSITET
USU UTAH STATE
Eau Claire
UNIVERSITY OF WISCONSIN STEVENS POINT
WISCONSIN STATE HERBARIUM UNIVERSITY OF WISCONSIN
WEST VIRGINIA UNIVERSITY
Yale

... and many more.

Data Ingestion Progress



November 2013 -

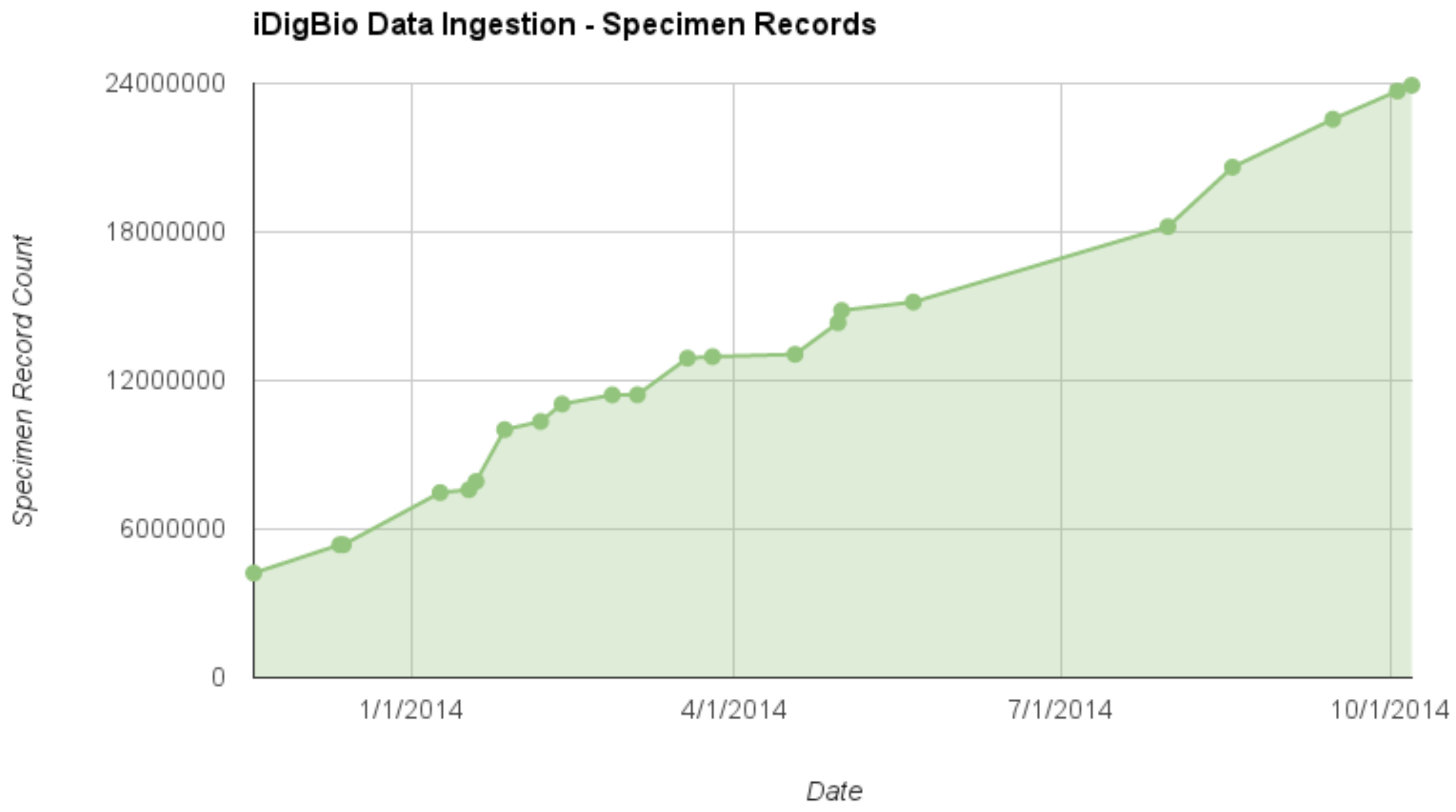
4.2 million specimen records

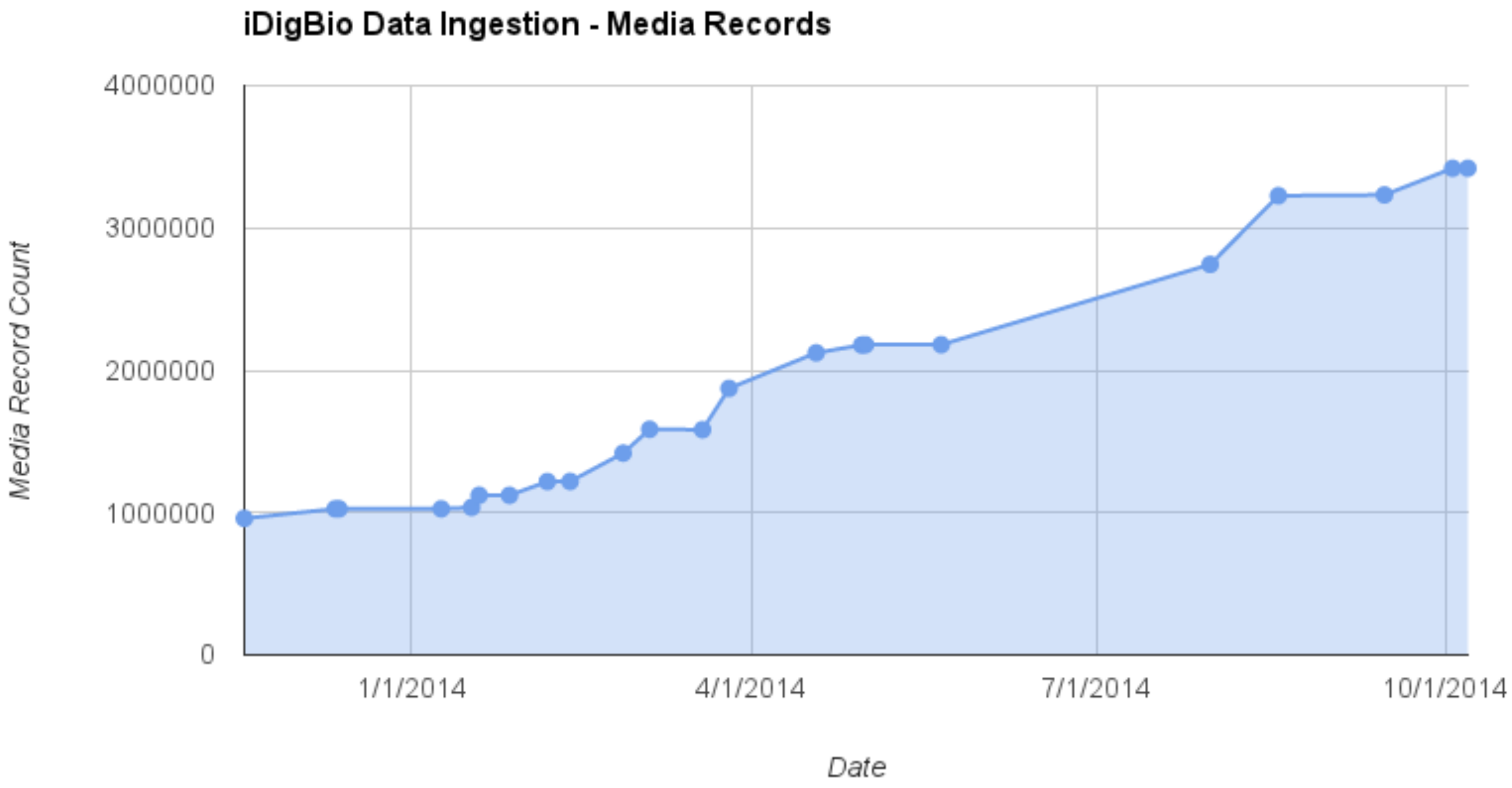
0.9 million media records

October 2014 -

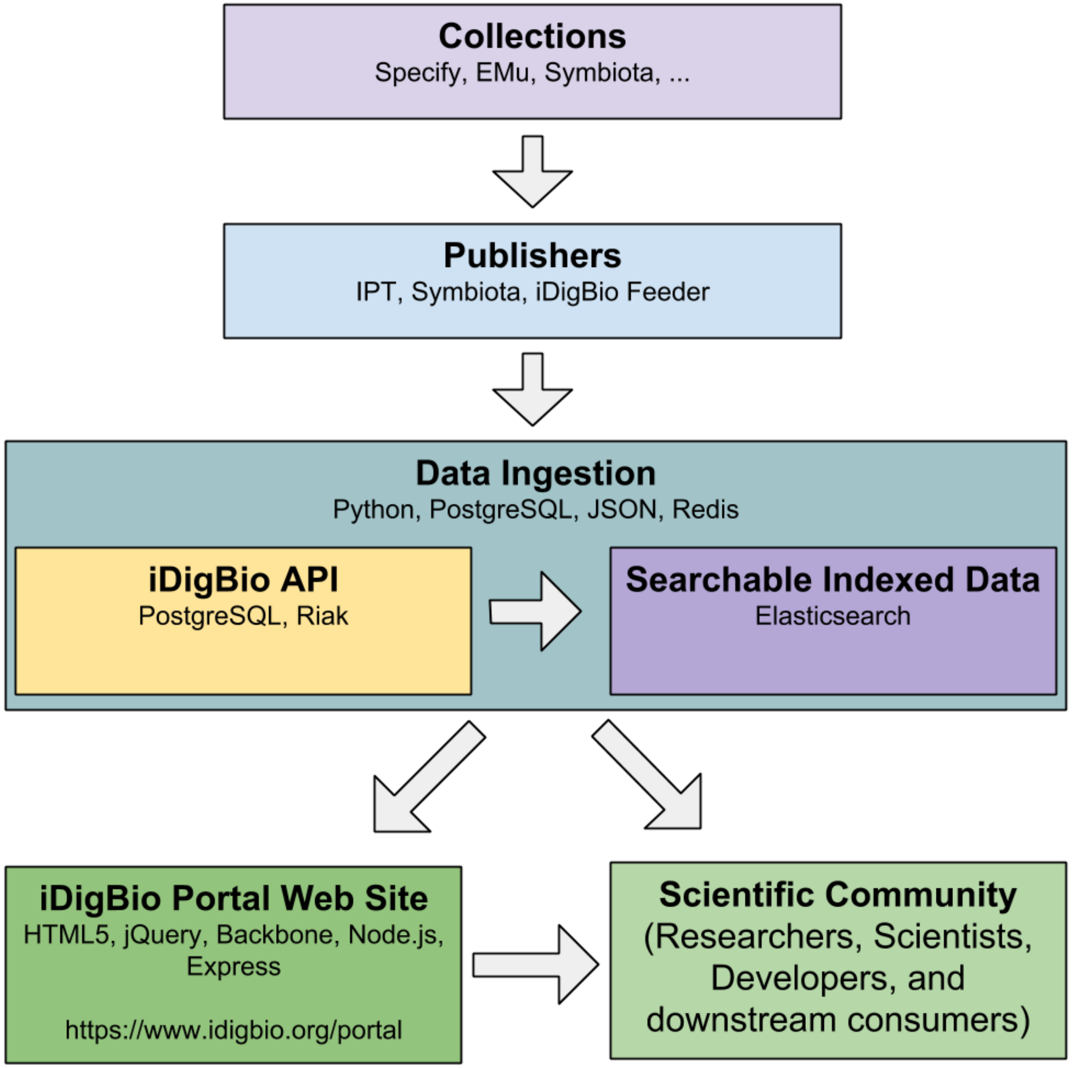
24 million specimen records

3.4 million media records


iDigBio Data Ingestion - Specimen Records
24000000
18000000
12000000
6000000
0
Specimen Record Count
1/1/2014
4/1/2014
7/1/2014
10/1/2014
Date


iDigBio Data Ingestion - Media Records
Media Record Count
4000000
3000000
2000000
1000000
0
1/1/2014
4/1/2014
7/1/2014
10/1/2014
Date

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)

Darwin Core Archive / DwC-A
http://rs.tdwg.org/dwc/terms/guides/text/

A Darwin Core Archive is a zip file that includes metadata about the dataset, the data itself, and any optional extension data.

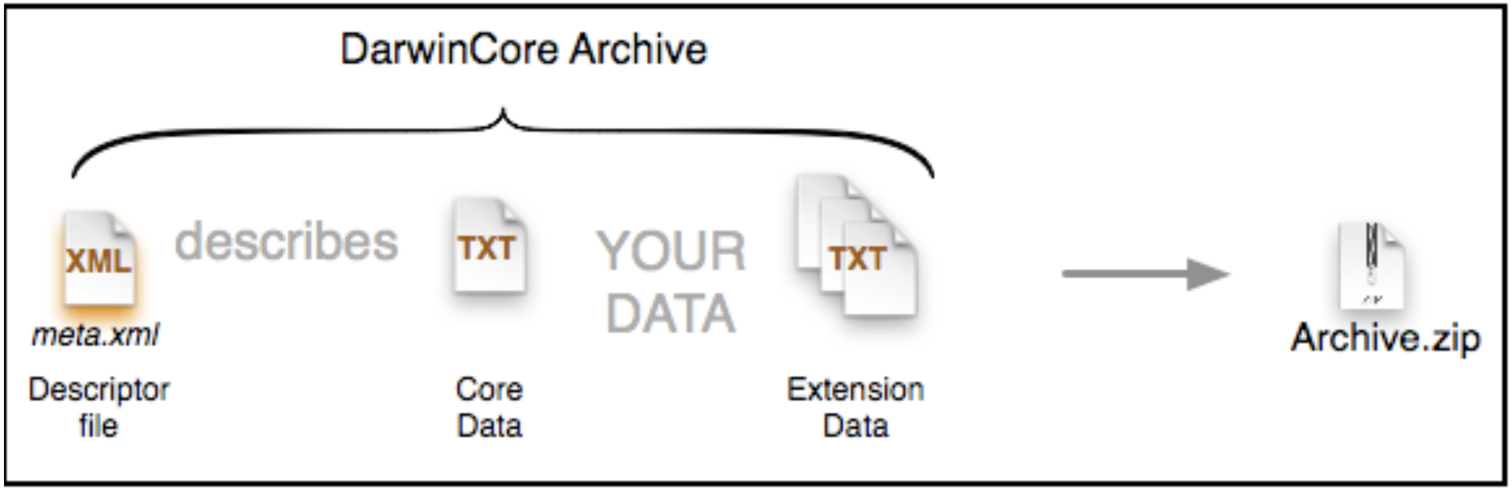
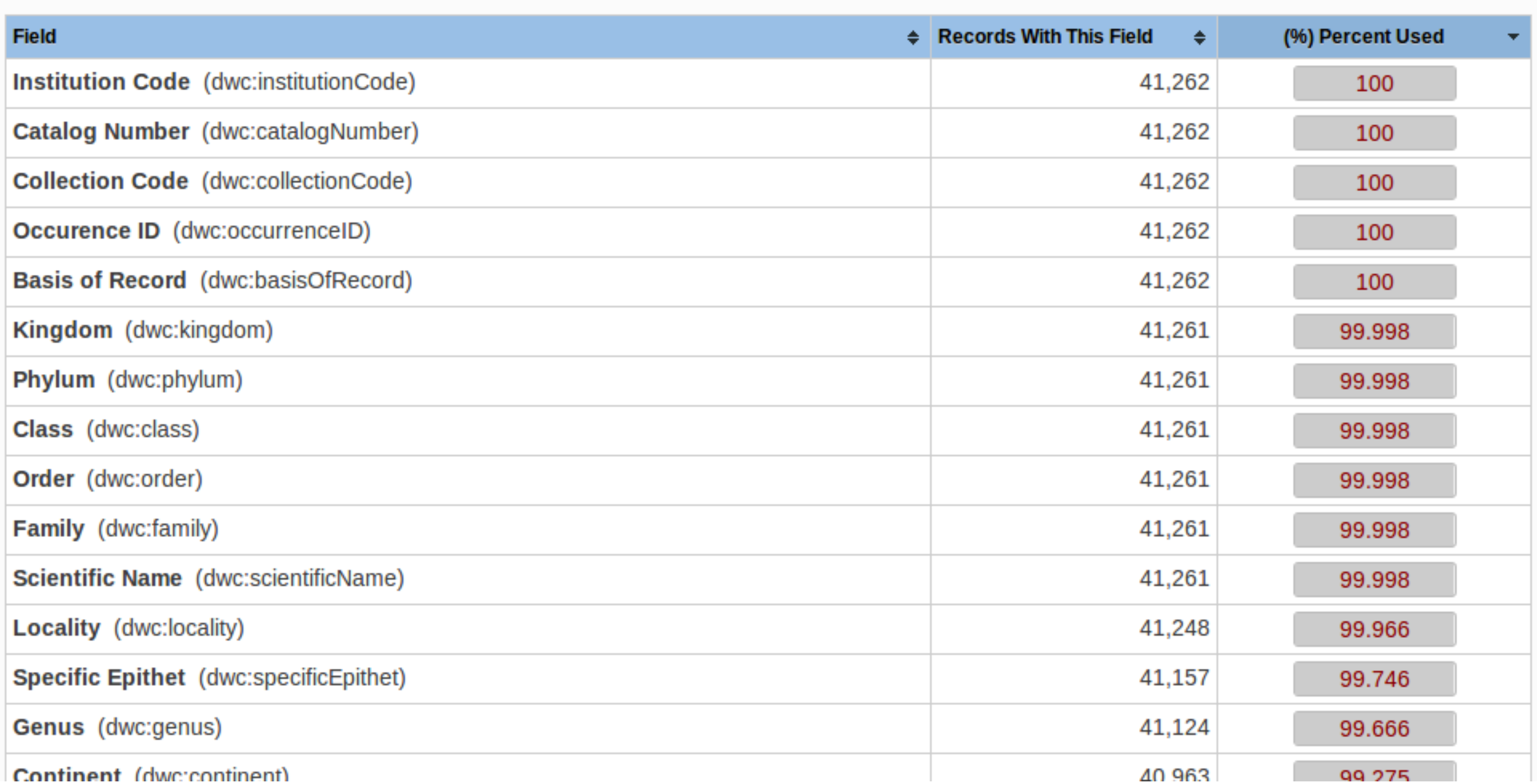


Image source: http://tools.gbif.org/dwca-assistant/

Specimen Data – Darwin Core Standard
http://rs.tdwg.org/dwc/terms/



Field	Records With This Field	(%) Percent Used
Institution Code (dwc:institutionCode)	41,262	100
Catalog Number (dwc:catalogNumber)	41,262	100
Collection Code (dwc:collectionCode)	41,262	100
Occurence ID (dwc:occurrenceID)	41,262	100
Basis of Record (dwc:basisOfRecord)	41,262	100
Kingdom (dwc:kingdom)	41,261	99.998
Phylum (dwc:phylum)	41,261	99.998
Class (dwc:class)	41,261	99.998
Order (dwc:order)	41,261	99.998
Family (dwc:family)	41,261	99.998
Scientific Name (dwc:scientificName)	41,261	99.998
Locality (dwc:locality)	41,248	99.966
Specific Epithet (dwc:specificEpithet)	41,157	99.746
Genus (dwc:genus)	41,124	99.666
Continent (dwc:continent)	40,963	99.275

Three types of data publishing technologies currently being consumed by iDigBio:

- **GBIF Integrated Publishing Toolkit (IPT)** - a Java tool used to publish and share biodiversity datasets http://www.gbif.org/ipt/

- **Symbiota** – web-based collection management software http://symbiota.org/

- **iDigBio RSS Feeder** – data sharing service for providers who do not run infrastructure

Data Source Types Providing Data to iDigBio

The following collection systems, databases, applications are known to have a capability to be a data source for iDigBio.

- Specify Software Project
- EMu Museum Management System
- Symbiota
- Arctos
- Excel
- …

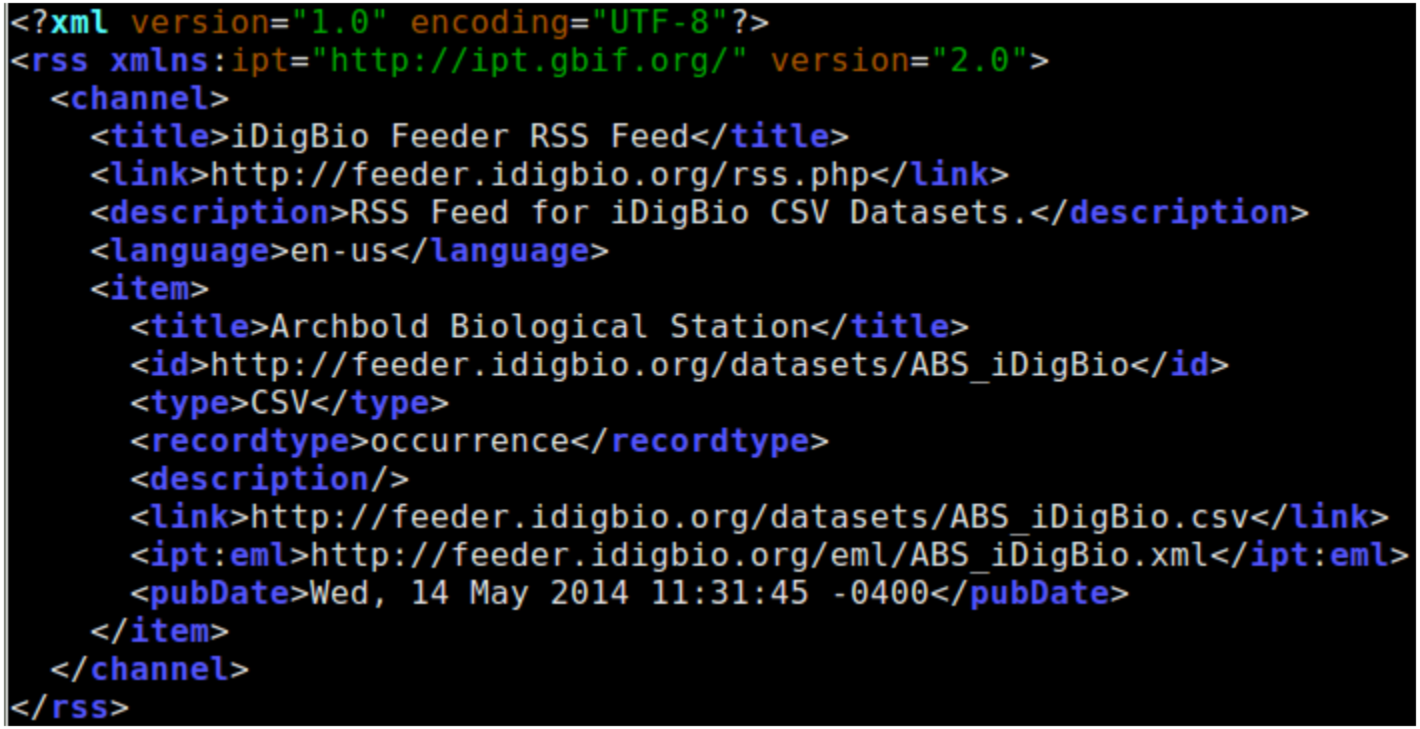
The iDigBio Mobilization Team (data@idigbio.org) assist with the preparation of data sets prior to Data Ingestion and are available to answer questions about sharing data with iDigBio.

See Also:

https://www.idigbio.org/wiki/index.php/Digitization_Resources

RSS is preferred format for Data Feeds

RSS (Really Simple Syndication) provides the mechanism to list available data files and share when they are updated. iDigbio reads the RSS feed to determine whether to download the data file again in order to collect updates.



```
<?xml version="1.0" encoding="UTF-8"?>
<rss xmlns:ipt="http://ipt.gbif.org/" version="2.0">
  <channel>
    <title>iDigBio Feeder RSS Feed</title>
    <link>http://feeder.idigbio.org/rss.php</link>
    <description>RSS Feed for iDigBio CSV Datasets.</description>
    <language>en-us</language>
    <item>
      <title>Archbold Biological Station</title>
      <id>http://feeder.idigbio.org/datasets/ABS_iDigBio</id>
      <type>CSV</type>
      <recordtype>occurrence</recordtype>
      <description/>
      <link>http://feeder.idigbio.org/datasets/ABS_iDigBio.csv</link>
      <ipt:eml>http://feeder.idigbio.org/eml/ABS_iDigBio.xml</ipt:eml>
      <pubDate>Wed, 14 May 2014 11:31:45 -0400</pubDate>
    </item>
  </channel>
</rss>
```

Recommended minimum fields for iDigBio Ingestion:

Record ID (recordId) - unique identifier for the digital record

Occurrence ID (occurenceID) - unique identifier for the physical object or establishment of an Occurrence

Scientific Name (scientificName) - the full scientific name

Event Date (eventDate) - date-time, preferably in ISO 8601

Collector (recordedBy) - collector name, number, or field number

Locality Data (...) - verbatim and decimal locality fields, continent, country, water body, state/province, ….

Catalog Number (catalogNumber) - Barcode, catalog number, accession id or collection number

Institution Code (institutionID) - institution identifier

Collection Code (collectionID) - collection identifier

Paleo specimens should also include Geological Context fields.

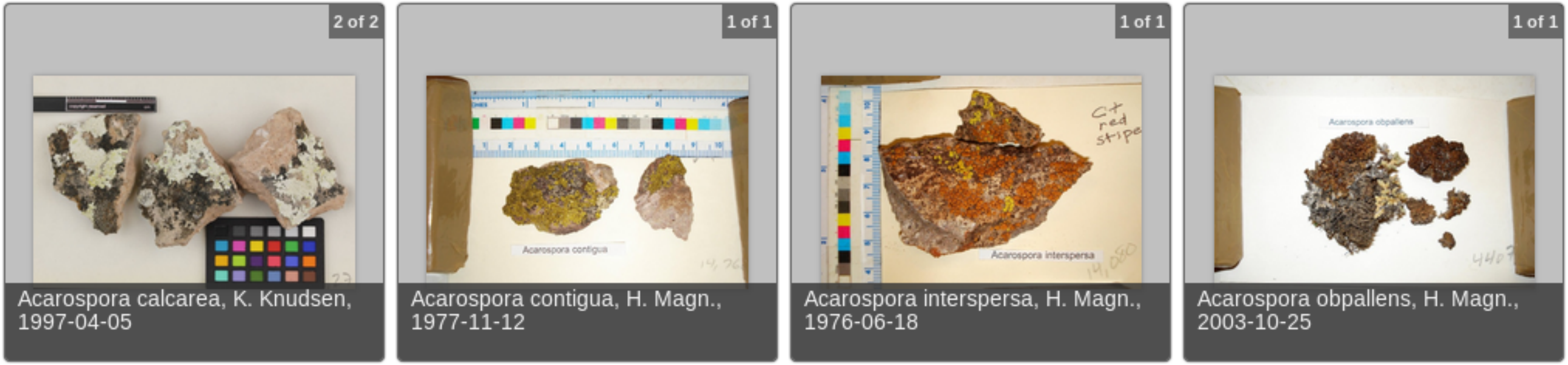
Dataset File Formats Consumable by iDigBio

- IPT – DwC-A
- Symbiota portals – DwC-A
- iDigBio Feeder – DwC-A, CSV, …

If you can export specimen data from your system / database / spreadsheet into DwC-A (or even CSV), then you can share data with iDigBio.

iDigBio RSS Feeder facilitates the sharing of over 1.5 million specimen records and 200 thousand media records from providers who do not need to run “servers”.

Media Data – Audubon Core / AC
http://terms.tdwg.org/wiki/Audubon_Core_Term_List



Acarospora calcarea, K. Knudsen, 1997-04-05

Acarospora contigua, H. Magn., 1977-11-12

Acarospora interspersa, H. Magn., 1976-06-18

Acarospora obpallens, H. Magn., 2003-10-25

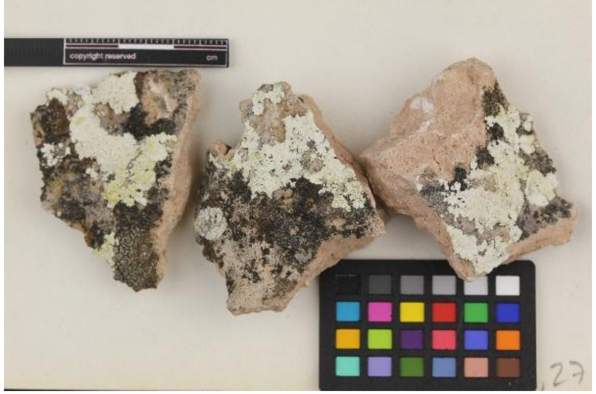
Images Source: Arizona State University Lichen Herbarium (Accessed through iDigBio Specimen Data Portal, https://www.idigbio.org/portal, 2014-09-18)

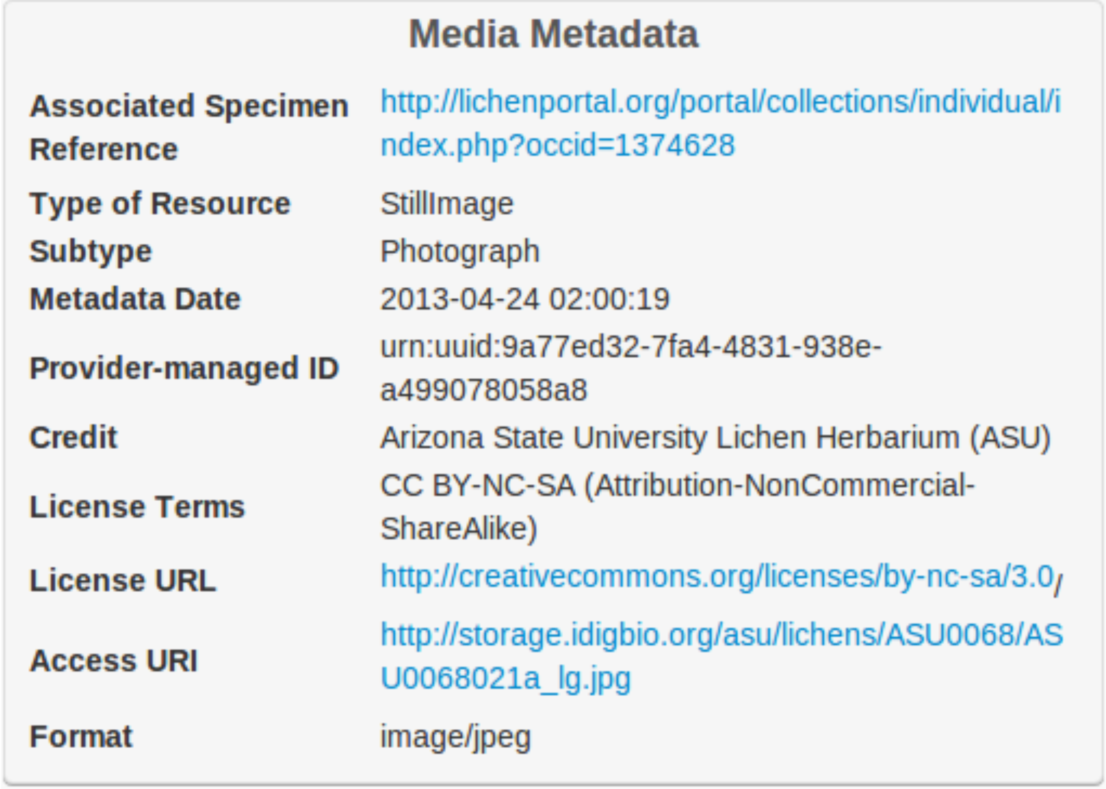
GBIF has a nice write-up on the benefits of AC over dwc:associatedMedia:

http://gbif.blogspot.com/2014/05/multimedia-in-gbif.html

Audubon Core vocabularies address such concerns as:

- the management of the media and collections
- descriptions of their content
- their taxonomic, geographic, and temporal coverage
- appropriate ways to retrieve, attribute and reproduce them





Media Metadata

Associated Specimen Reference	http://lichenportal.org/portal/collections/individual/index.php?occid=1374628
Type of Resource	StillImage
Subtype	Photograph
Metadata Date	2013-04-24 02:00:19
Provider-managed ID	urn:uuid:9a77ed32-7fa4-4831-938e-a499078058a8
Credit	Arizona State University Lichen Herbarium (ASU)
License Terms	CC BY-NC-SA (Attribution-NonCommercial-ShareAlike)
License URL	http://creativecommons.org/licenses/by-nc-sa/3.0/
Access URI	http://storage.idigbio.org/asu/lichens/ASU0068/ASU0068021a_lg.jpg
Format	image/jpeg

Audubon Core can support “new” media types

Specimen Data

dwc:catalogNumber: UF 105199

dwc:scientificName:

Carcharocles megalodon

dwc:stateProvince: Florida

dwc:county: Duval

dwc:latestPeriodOrHighestSystem:

Late Miocene

dwc:decimalLatitude: 30.39211

Media Data

dwc:scientificName:

Carcharocles megalodon

dc:type: image

ac:subtype: http://www.fabbers.com/StL.asp

ac:subtypeLiteral: 3dModel

ac:tag: tooth

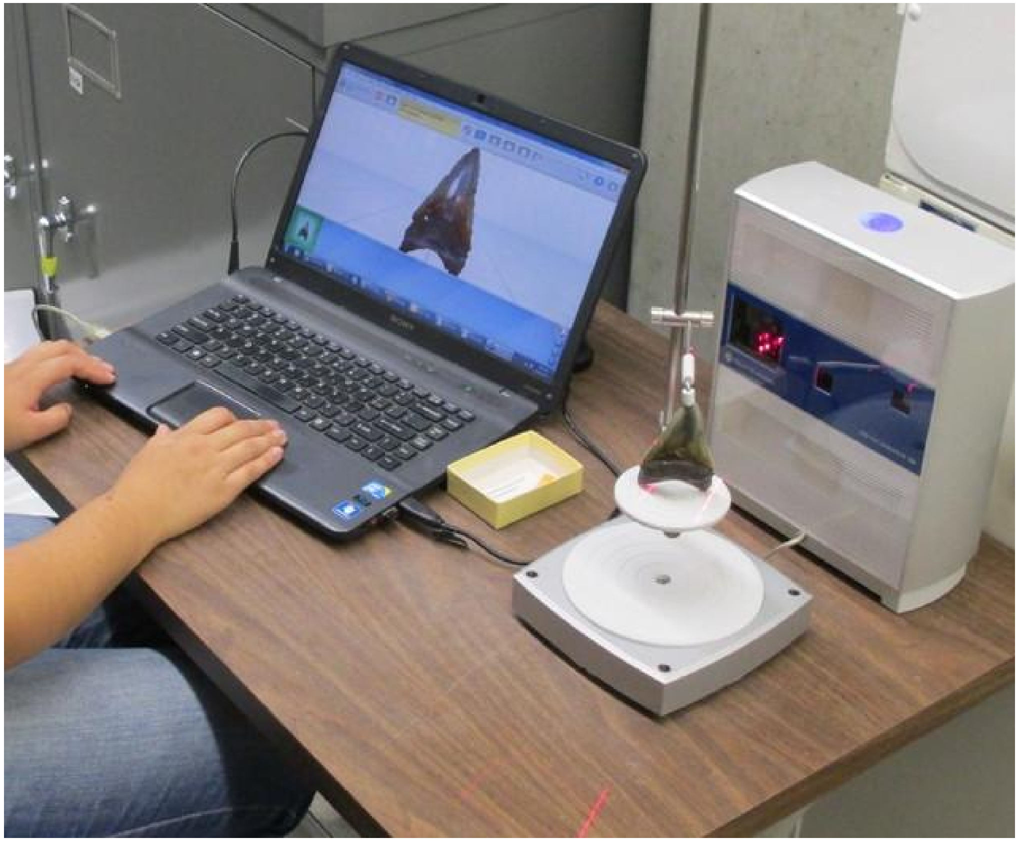


Image source: Aaron Wood, Florida Museum of Natural History

Darwin Core Archive - Extension to link data between the Occurrence and Audubon Core record

Specimen Data

dwc:occurrenceID: **3bca767a-5a25-42c0-12...**

dwc:scientificName: Carcharocles megalodon

dwc:catalogNumber: UF 105199

dwc:stateProvince: Florida

dwc:county: Duval

dwc:latestPeriodOrHighestSystem: Miocene

dwc:decimalLatitude: 30.39211

...

Media Data

dcterms:identifier: 9a3025b1-f686-4e43-915f-...

coreid: **3bca767a-5a25-42c0-12...**

dwc:scientificName: Carcharocles megalodon

ac:associatedSpecimenReference: http://museum...

dc:type: image

ac:subtype: http://www.fabbers.com/StL.asp

ac:subtypeLiteral: 3dModel

ac:tag: tooth

...

In the wild, ac:associatedSpecimenReference tends NOT to provide the bare occurrence id of the related specimen, so instead we use the implicit relationship via coreid in the DwC-A.

Image source: http://commons.wikimedia.org/wiki/File:Carcharocles_megalodon_tooth.JPG

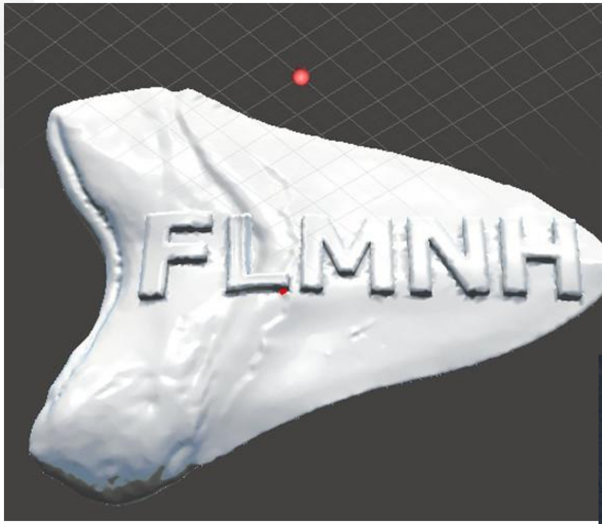


Image source: Aaron Wood

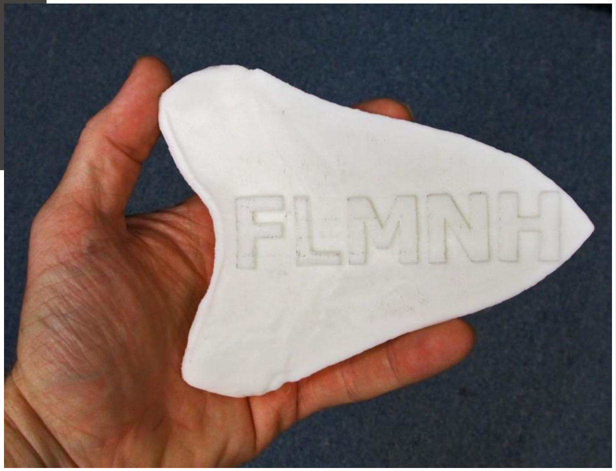


Image source: Aaron Wood
3D Model printing by Robert Burns

Recommended minimum Audubon Core fields for iDigBio Data Ingestion:

Access URI
Rights
Provider
Scientific name
Title
Description
Tags

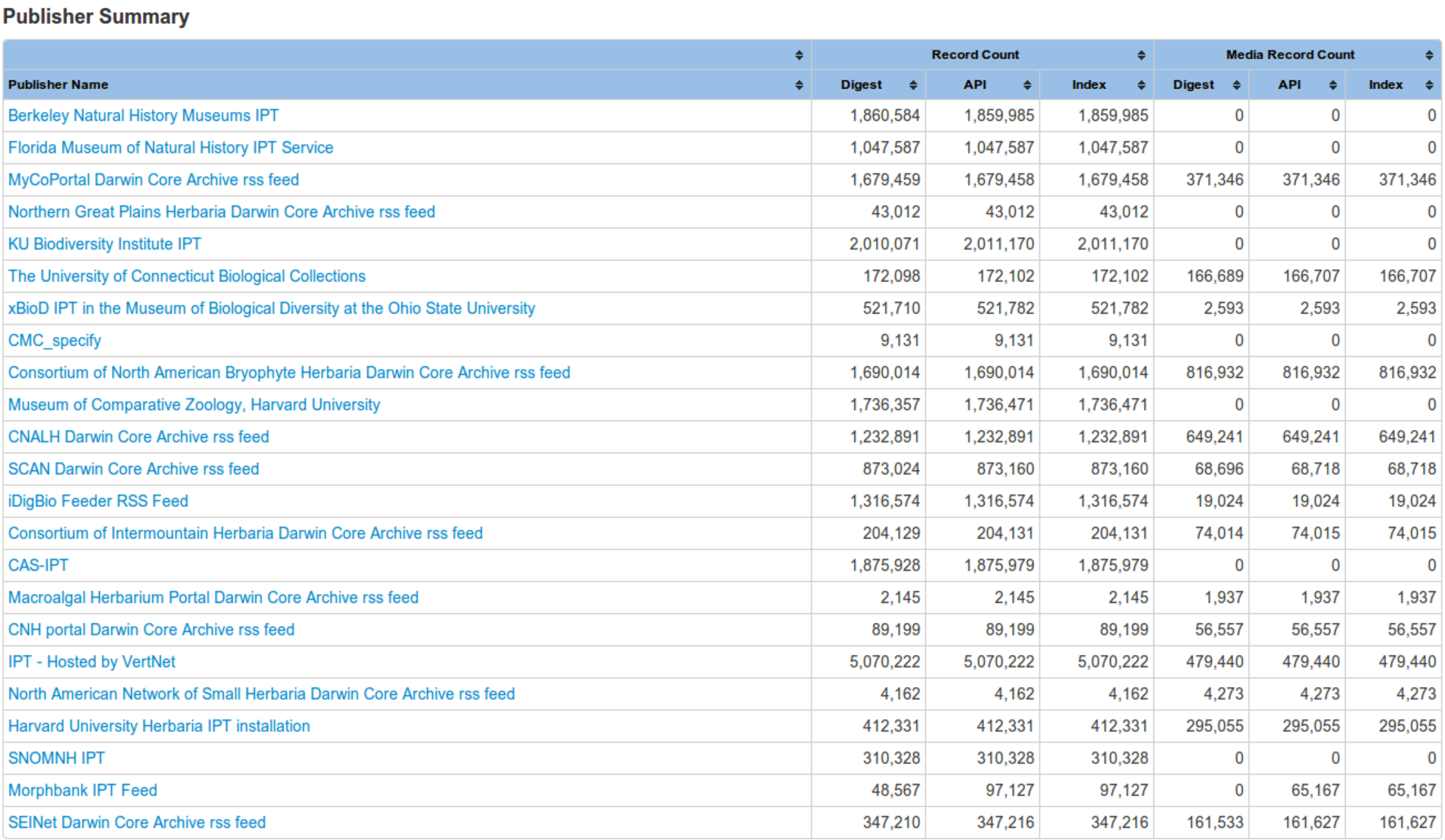
Practical Details

- Data Formats
  - ISO 8601 Dates
  - WGS84 Decimal Lat/Long
- Controlled Vocabularies
  - ISO Country Names and Codes
  - State/Province names
- Identifier Formats (UUID, ARK, URN, DOI, URI, URL, LSID, ...)
- Copyright and Standard Licenses
- Apple Core guidelines for herbaria
  - http://code.google.com/p/applecore/wiki/Introduction

Ingestion Process Changes Over the Past Year

- New Staff (Dan Stoner… that’s me!)
- Improved parallelization of ingestion tasks
- Incremental Indexing
- Database Tuning
- Ingestion Reporting

Ingestion Reporting
https://www.idigbio.org/portal/publishers



Publisher Summary

	Record Count			Media Record Count		
Publisher Name	**Digest**	**API**	**Index**	**Digest**	**API**	**Index**
Berkeley Natural History Museums IPT	1,860,584	1,859,985	1,859,985	0	0	0
Florida Museum of Natural History IPT Service	1,047,587	1,047,587	1,047,587	0	0	0
MyCoPortal Darwin Core Archive rss feed	1,679,459	1,679,458	1,679,458	371,346	371,346	371,346
Northern Great Plains Herbaria Darwin Core Archive rss feed	43,012	43,012	43,012	0	0	0
KU Biodiversity Institute IPT	2,010,071	2,011,170	2,011,170	0	0	0
The University of Connecticut Biological Collections	172,098	172,102	172,102	166,689	166,707	166,707
xBioD IPT in the Museum of Biological Diversity at the Ohio State University	521,710	521,782	521,782	2,593	2,593	2,593
CMC_specify	9,131	9,131	9,131	0	0	0
Consortium of North American Bryophyte Herbaria Darwin Core Archive rss feed	1,690,014	1,690,014	1,690,014	816,932	816,932	816,932
Museum of Comparative Zoology, Harvard University	1,736,357	1,736,471	1,736,471	0	0	0
CNALH Darwin Core Archive rss feed	1,232,891	1,232,891	1,232,891	649,241	649,241	649,241
SCAN Darwin Core Archive rss feed	873,024	873,160	873,160	68,696	68,718	68,718
iDigBio Feeder RSS Feed	1,316,574	1,316,574	1,316,574	19,024	19,024	19,024
Consortium of Intermountain Herbaria Darwin Core Archive rss feed	204,129	204,131	204,131	74,014	74,015	74,015
CAS-IPT	1,875,928	1,875,979	1,875,979	0	0	0
Macroalgal Herbarium Portal Darwin Core Archive rss feed	2,145	2,145	2,145	1,937	1,937	1,937
CNH portal Darwin Core Archive rss feed	89,199	89,199	89,199	56,557	56,557	56,557
IPT - Hosted by VertNet	5,070,222	5,070,222	5,070,222	479,440	479,440	479,440
North American Network of Small Herbaria Darwin Core Archive rss feed	4,162	4,162	4,162	4,273	4,273	4,273
Harvard University Herbaria IPT installation	412,331	412,331	412,331	295,055	295,055	295,055
SNOMNH IPT	310,328	310,328	310,328	0	0	0
Morphbank IPT Feed	48,567	97,127	97,127	0	65,167	65,167
SEINet Darwin Core Archive rss feed	347,210	347,216	347,216	161,533	161,627	161,627

Planned Future Changes

- Parallelize more parts of Ingestion process (such as media processing)
- Support for additional publisher types (beyond IPT, Symbiota, iDigBio RSS Feeder)
- Improved Ingestion logging and error detection
- Support for additional media types (audio, 3D scans, …)
- Data Quality

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. All images used with permission or are free from copyright.

End.