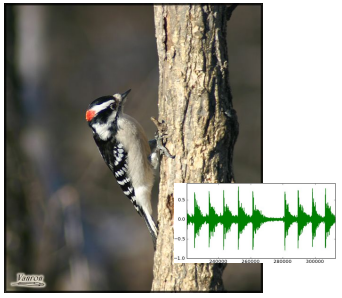
Assuring the quality of your data: A natural history collection community perspective

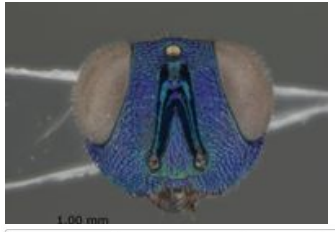
Deborah Paul, Katja Seltmann, Laura Russell, David Bloom
@idbdeb @iDigBio @VertNetOrg

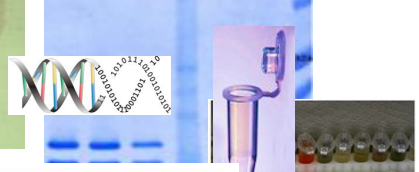
Tuesday, January 12, 2016
9 am Pacific / 10 am Mountain /
11 am Central / 12 noon Eastern

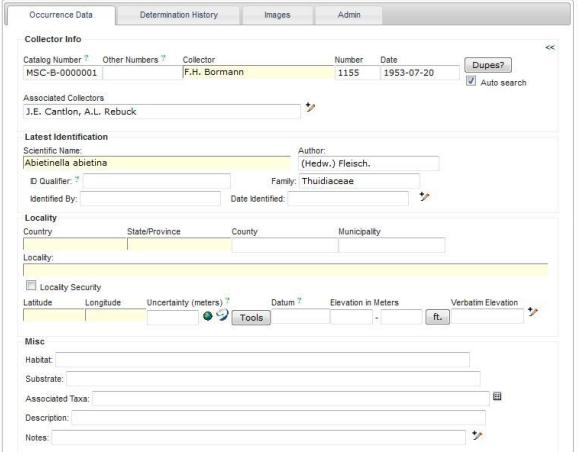

20 mm

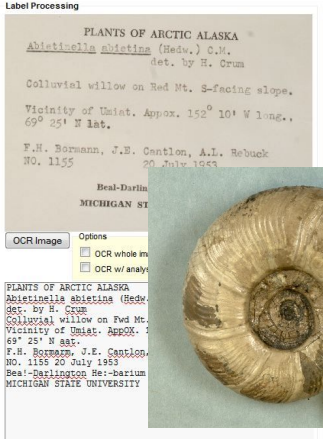



R.K. GODFREY HERBARIUM
202#3
FLORIDA STATE UNIVERSITY
PLANTS OF FLORIDA, USA

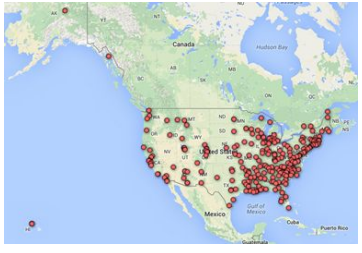

1.00 mm




Occurrence Data
Determination History
Images
Admin
Collector Info
Catalog Number
Other Numbers
Collector
Number
Date
MSC-B-0000001
F.H. Bormann
1155
1953-07-20
Dupes?
Auto search
Associated Collectors
J.E. Cantlon, A.L. Rebuck
Latest Identification
Scientific Name:
Author:
Abietinella abietina
(Hedw.) Fleisch.
ID Qualifier:
Family:
Thuidiaceae
Identified By:
Date Identified:
Locality
Country
State/Province
County
Municipality
Locality:
Locality Security
Latitude
Longitude
Uncertainty (meters)
Datum
Elevation in Meters
Verbatim Elevation
Tools
ft.
Misc
Habitat:
Substrate:
Associated Taxa:
Description:
Notes:

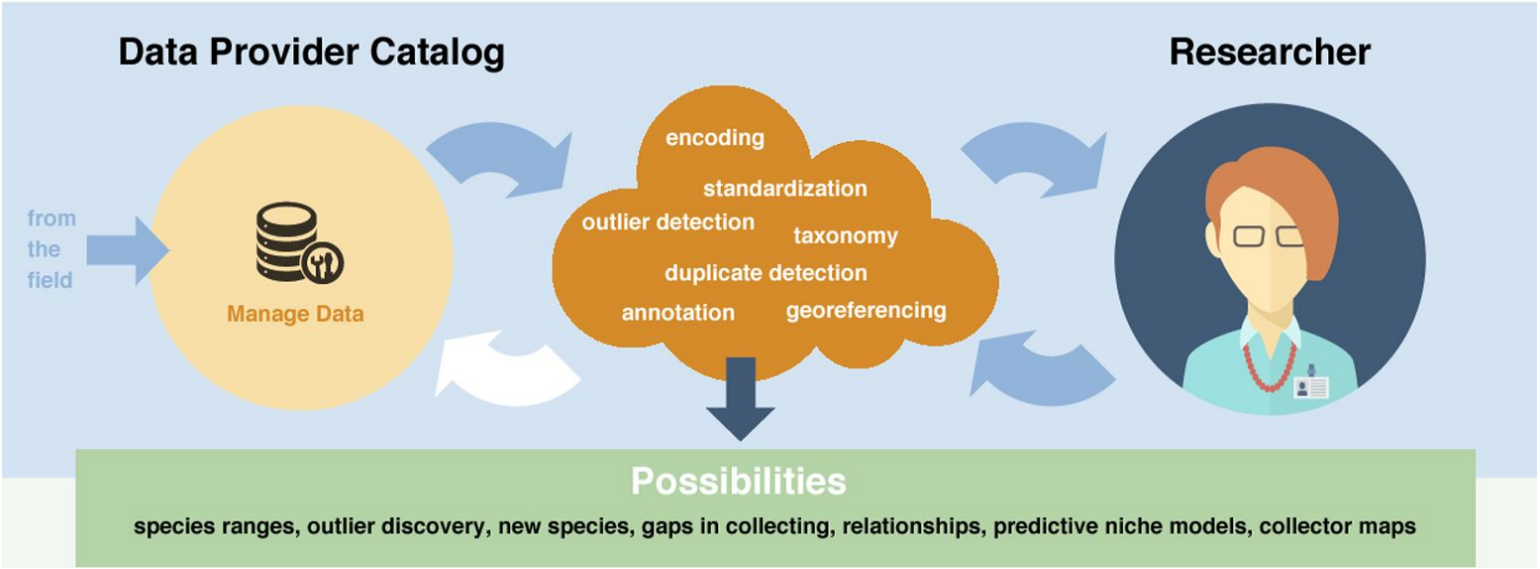

Label Processing
PLANTS OF ARCTIC ALASKA
Abietinella abietina (Hedw.) C.M.
det. by H. Crum
Colluvial willow on Red Mt. S-facing slope.
Vicinity of Umiat. Appox. 152° 10' W long., 69° 25' N lat.
F.H. Bormann, J.E. Cantlon, A.L. Rebuck
NO. 1155
20 July 1953
Beal-Darlington Herbarium
MICHIGAN STATE UNIVERSITY
OCR Image
Options
OCR whole im
OCR w/ analys


NAME
MILTON S. RAY COLLECTION
DEPT. NO.
Procephalus Grundiaceus
LOCALITY
COLLECTOR
Prussia, Germany


Canada
United States
Mexico
Gulf of Mexico
Cuba
Puerto Rico

Our data producers & data users

Data Quality starts here, before collection of specimens and field data



Naturalists, Field biologists,
Nature explorers, Research institutions, Citizen scientists, Curators

Ecologists, biogeographers; Analysts, modellers;
Conservation planners;
Nature managers; Policy managers;
Funding agencies; Industry; ?, ...

Data, data types, data standards

Darwin Core (DwC)
Audubon Media (AC)
Ecological Metadata Language (EML)
Global *Genome* Biodiversity Network (*GGBN*)
, • • •

DwC Categories

- **Record Level** (19)
- Occurrence (19)
- Organism (7)
- Material Sample (1)
- Event (15)
- Location (44)
- Geological Context (18)
- Identification (8)
- Taxon (33)

http://rs.tdwg.org/dwc/terms/index.htm

Record Level

dcterms:type | dcterms:modified | **dcterms:language | dcterms:license** | **dcterms:rightsHolder** | **dcterms: accessRights** | dcterms: bibliographicCitation | dcterms: references | **institutionID** | **collectionID** | datasetID | **institutionCode** | **collectionCode** | datasetName | ownerInstitutionCode | **basisOfRecord** | **informationWithheld** | dataGeneralizations | dynamicProperties

Data Publication Fallacies (and Truths)

The Fallacy of Perfection:

Data must be perfect before publication.

The Truth:

Data will never be perfect in every aspect.

Data Publication Fallacies (and Truths)

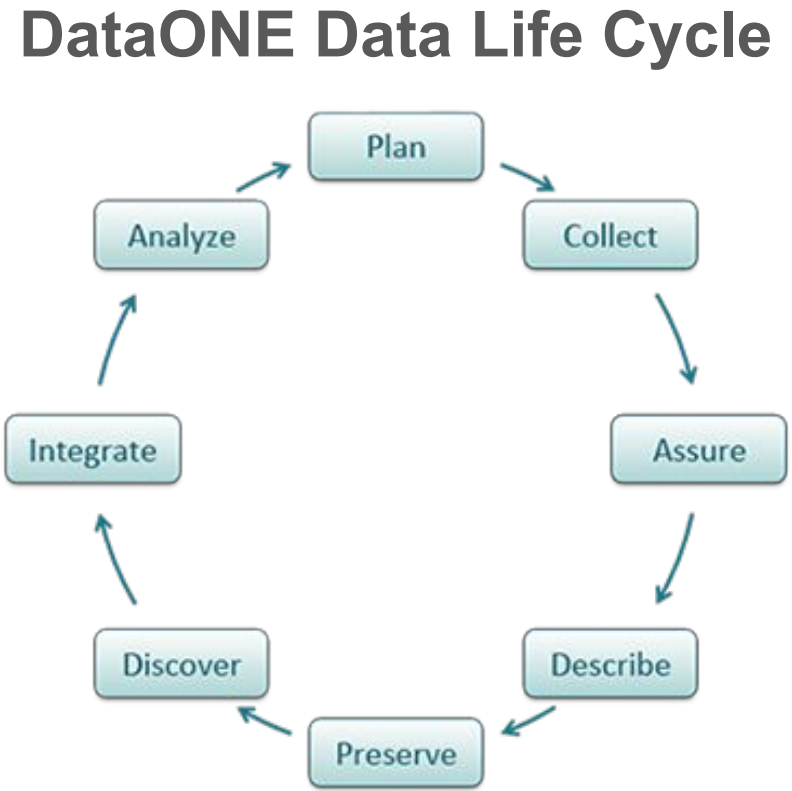
The Fallacy of Petrification:

Data do not change (or don’t need to change) once they are in a ledger or database.

The Truth:

Data are dynamic and require regular curation.

Data Publication Fallacies (and Truths)



Data Publication Fallacies (and Truths)

The Fallacy of Fitness for Use:

The fitness for use of data depends upon how and why the data were collected.

The Truth:

Fitness depends upon the questions being asked (value is is in the eye of the beholder).

Challenges for data publication and quality

Individual:

Education, Experience, Funding, Presence of support

Institutional:

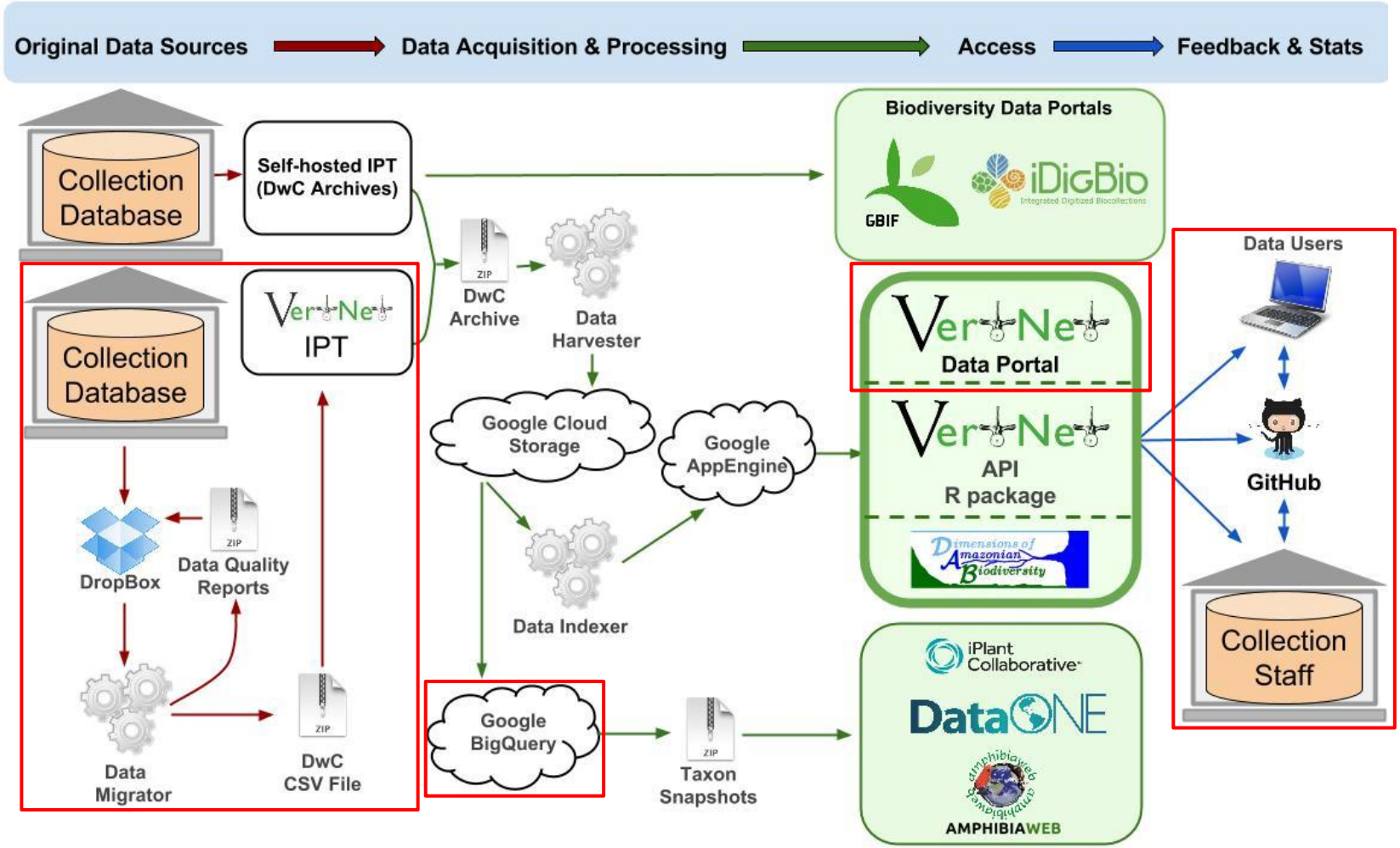
Inter/Intra-collection communication, Mixed technology, Legal limitations

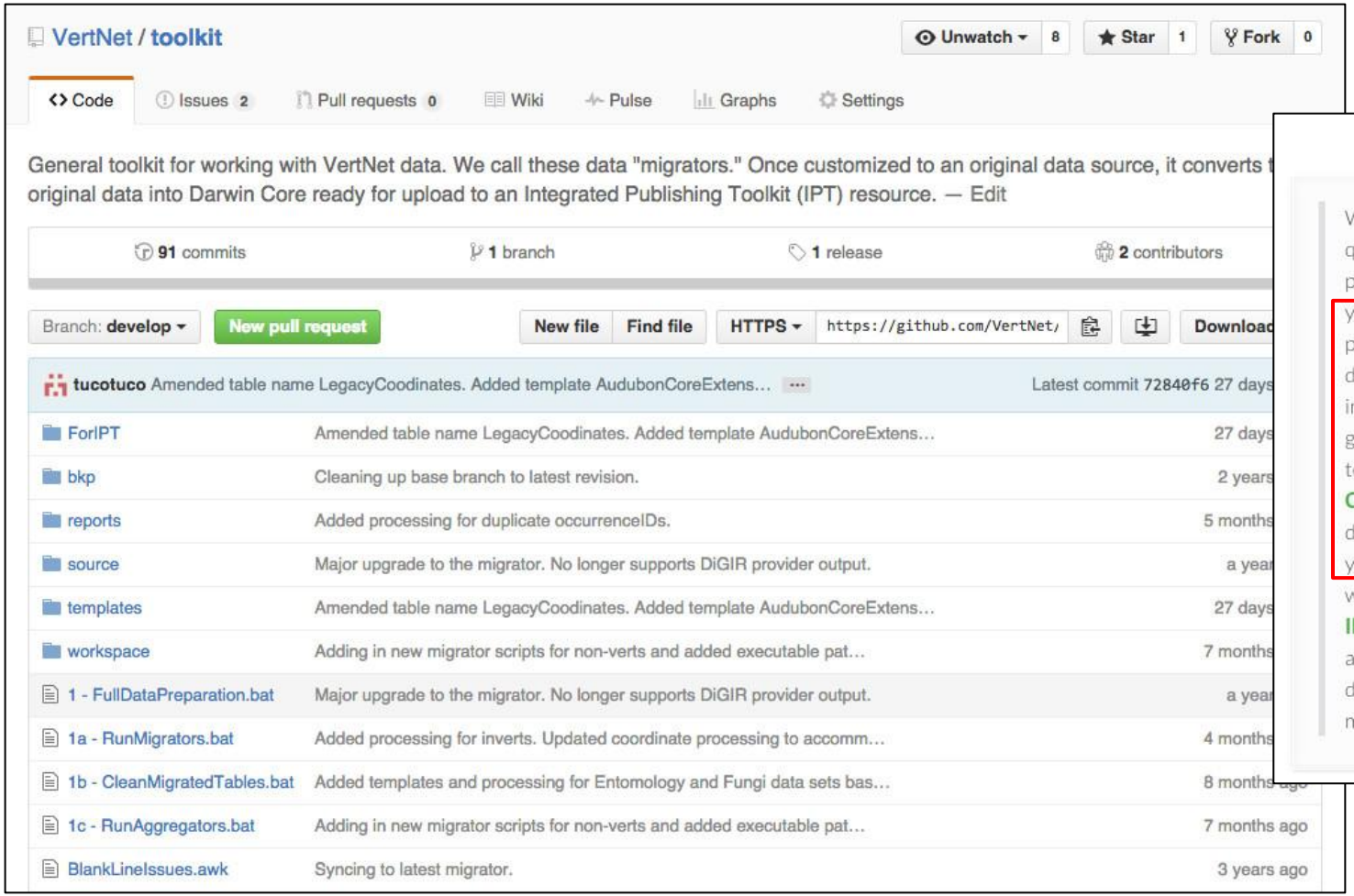
Structural:

Interoperability, Related/Derivative sources

Data assurance practices - VertNet

- Data publishing
  - VertNet Toolkit
    - Data migrator
    - Data quality reports
- Data acquisition and processing
  - Google BigQuery
- Data access - VertNet portal
  - Spatial quality
  - Flag data issues
- Feedback
  - Issue tracking and feedback via GitHub


Original Data Sources
Data Acquisition & Processing
Access
Feedback & Stats
Collection Database
Self-hosted IPT (DwC Archives)
Biodiversity Data Portals
GBIF
iDigBio
Integrated Digitized Biocollections
Collection Database
VertNet IPT
DwC Archive
Data Harvester
VertNet Data Portal
Data Users
Google Cloud Storage
Google AppEngine
VertNet API R package
Dimensions of Amazonian Biodiversity
GitHub
DropBox
Data Quality Reports
Data Indexer
Collection Staff
Data Migrator
DwC CSV File
Google BigQuery
Taxon Snapshots
iPlant Collaborative
DataONE
AMPHIBIAWEB



DATA QUALITY

We believe in publishing the highest quality and most complete data possible. After we talk with you about your data, we'll review your data for possible data improvements, such as duplicate catalog numbers, indeterminate and non-standard geography, inconsistent taxonomy, and terms not compliant with **Darwin Core**. Then we'll provide you with a detailed **report** so that you can update your database locally. Only publishers who host their data on the **VertNet IPT** benefit from these services automatically. If you host your own data set, you can request this service to make your data more complete.

https://github.com/VertNet/toolkit

Google BigQuery

New Query

```
1 SELECT class FROM [dumps.vertnet_latest] where class is not null group by class
```

RUN QUERY | Save Query | Save View | Format Query | Show Options | Query complete (4.1s

Results | Explanation

Row	class
155	Insecta
156	LEPOSPONDYL
157	LISSAMPHIBIAN
158	Leptocardii
159	Liliopsida
160	Lycopodiopsida
161	MAMM INSECT TRACE
162	MAMM INVERT
163	MAMM PLANT
164	MAMM TRACE
165	MAMMALIA

NCSM NCSM-Fishes 12174 *Oncorhynchus mykiss*

Submit data issue

Darwin Core | Rights | Spatial quality | Index

Warning: Some validations could not be performed. Check below.

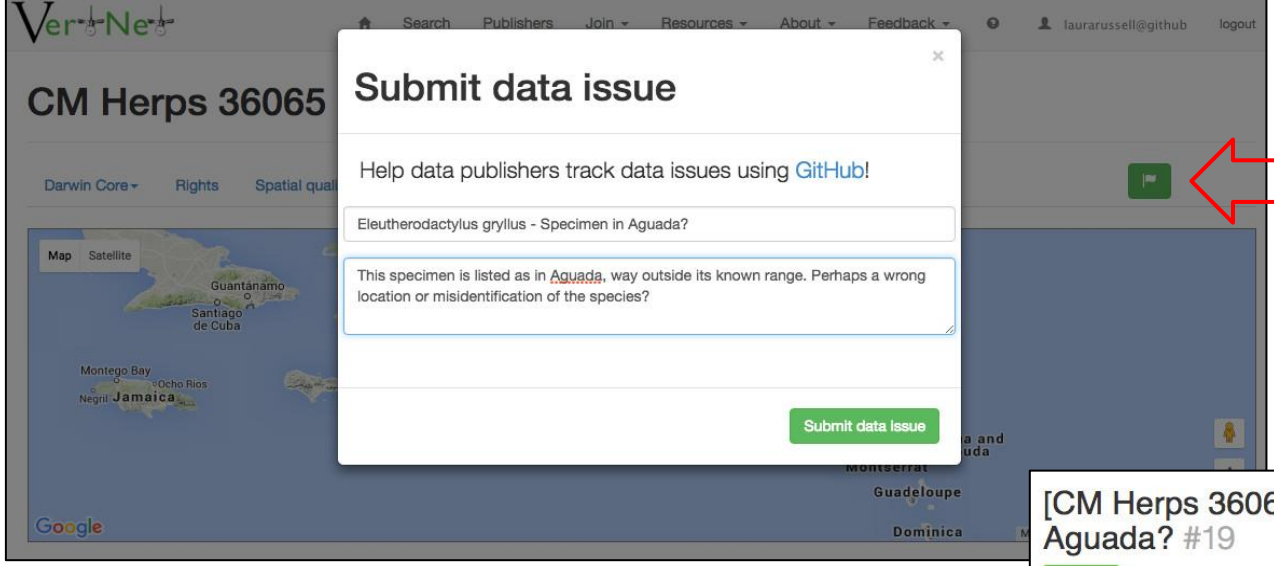
Data completeness		
Are coordinates present?	✔	Yes
Is the country value present?	✔	Yes
Are both coordinates 0 (zero)?	✔	No
Do coordinates have three or more decimal figures?	✔	Yes
Do coordinates have datum?	✔	Yes

Data inconsistencies		
Are coordinates within specified country?[1]	✔	Yes
Distance outside of specified country (in degrees)[1]	✔	0
Distance outside of species range map (in degrees)[1]	⊘	Could not be assessed

Data Errors		
Is latitude between 90 and -90?	✔	Yes
Is longitude between 180 and -180?	✔	Yes
Are coordinates transposed?[1]	✔	No
Is latitude hemisphere correct?[1]	✔	Yes
Is longitude hemisphere correct?[1]	✔	Yes

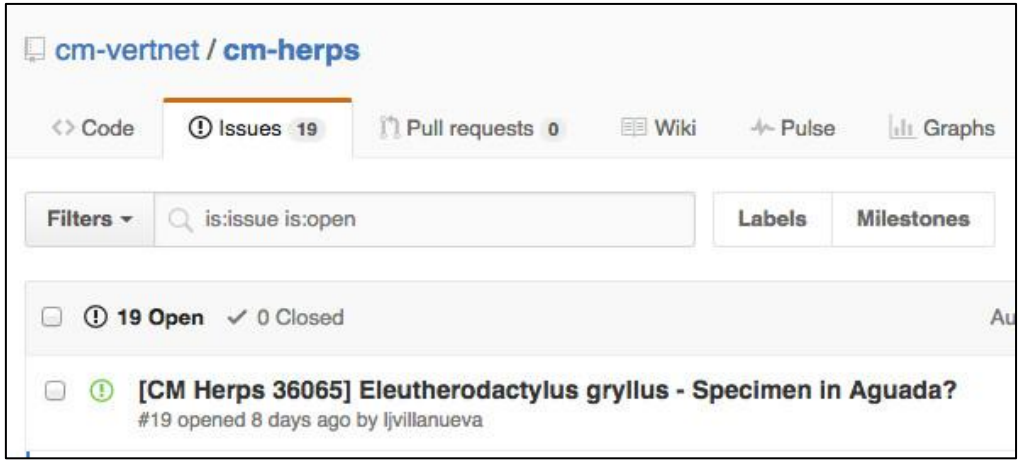
[1]*Assessed with Map Of Life validation tools*

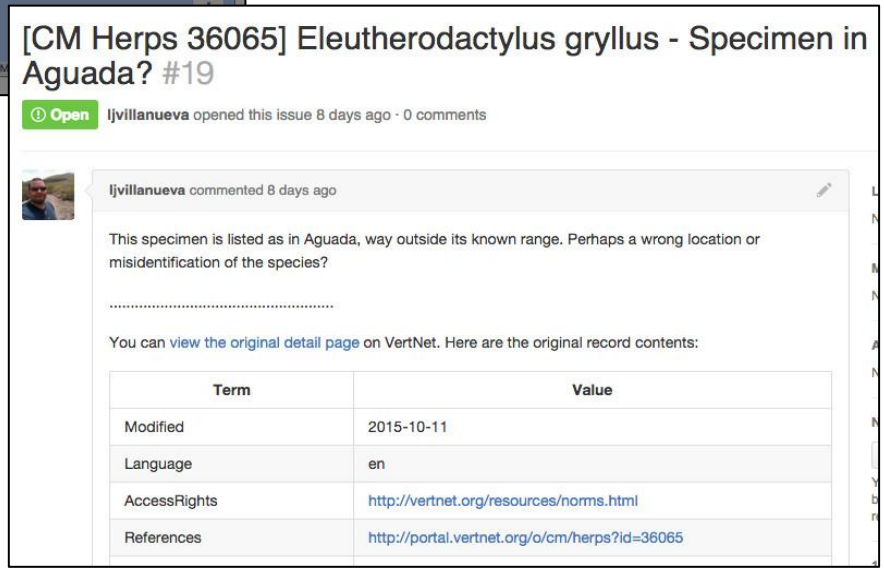
http://vertnet.org/resources/spatialqualitytabguide.html



Easy for users to submit. Users do need a free GitHub account.

Easy for data publishers to receive and manage feedback about their published data.





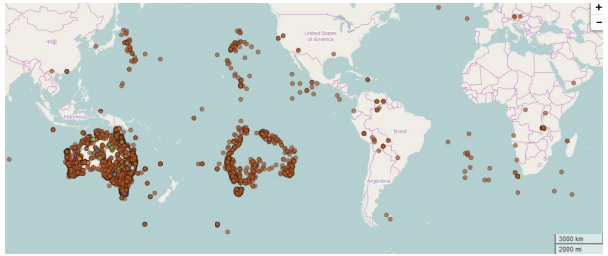
http://vertnet.org/resources/issuetrackingguide.html

Data assurance practices

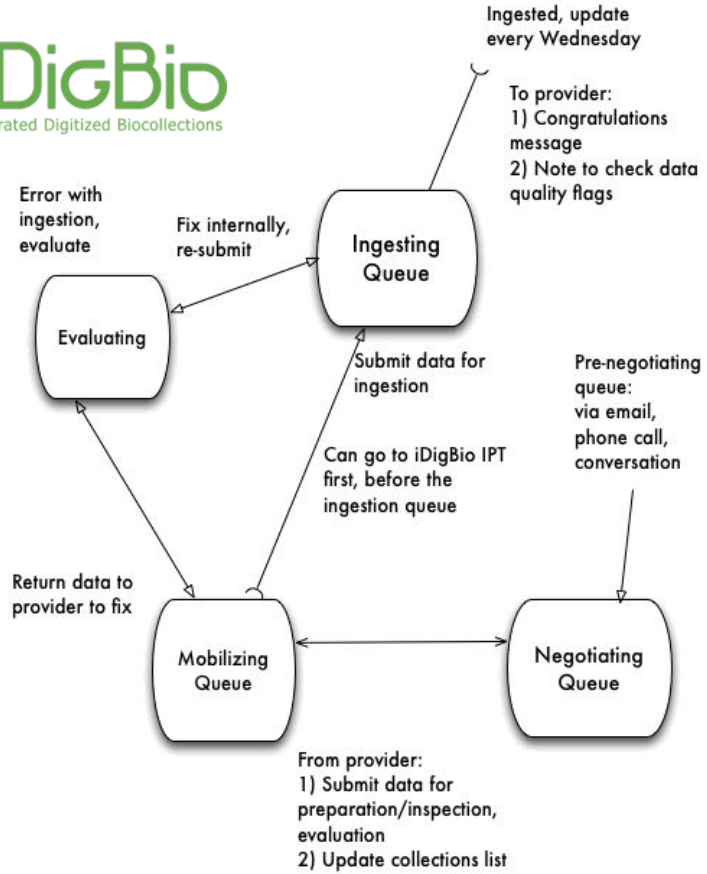
Data Publication - an overview of the process data goes through before ingestion at iDigBio

1. Negotiating
2. Mobilizing (human, scripts)
3. Ingestion
    a. Evaluation (human, scripts, reports)
    b. *iDigBio Data Quality (DQ) Flags*

a Perk of data sharing!



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



These queue name roughly line up with the ingestion report here: https://www.idigbio.org/wiki/index.php/Data_Ingestion_Report

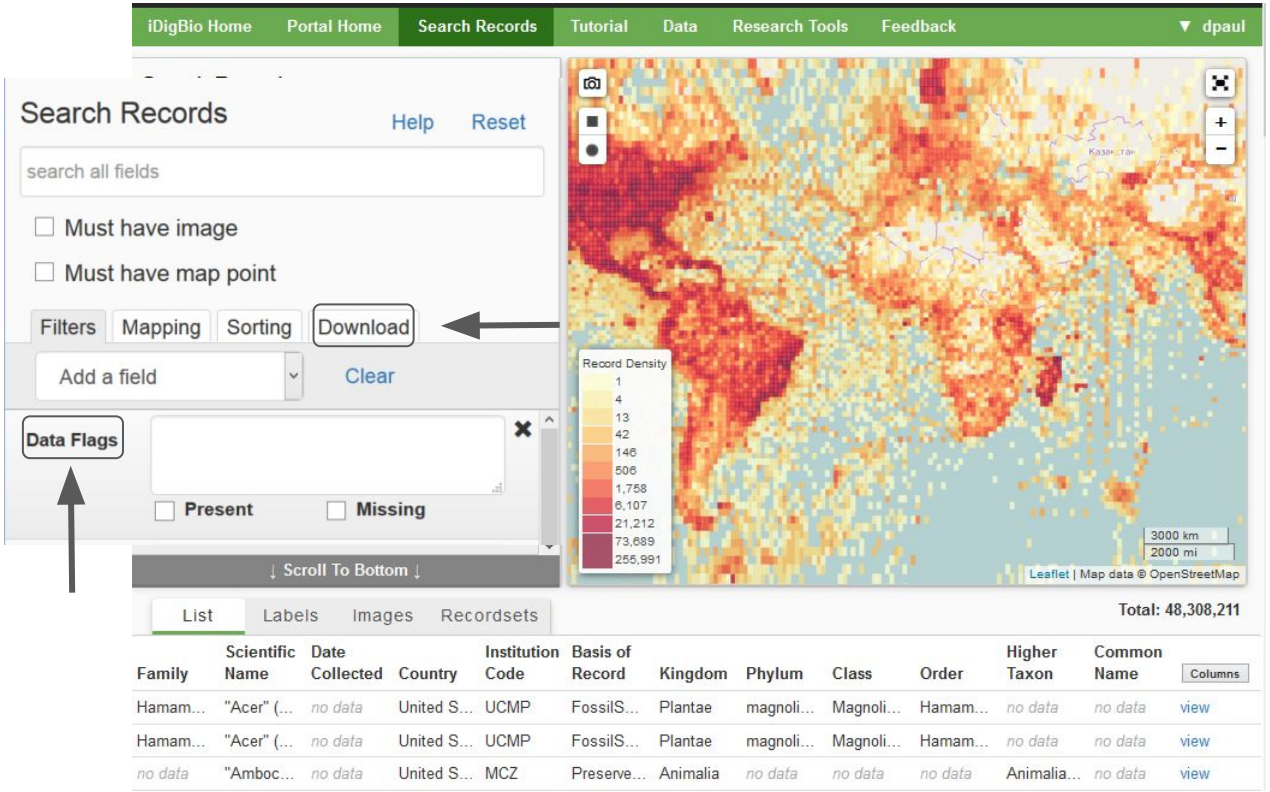
Graphic by Joanna McCaffrey, iDigBio Biodiversity Informatics Manager

https://www.idigbio.org/portal/search

Search and download using the iDigBio DQ Flags

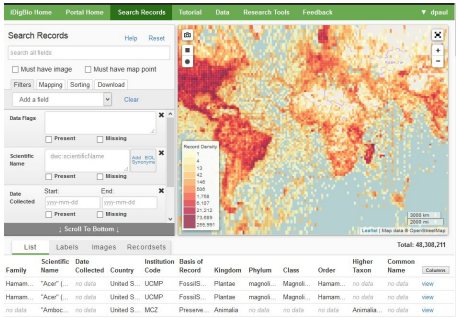
DQ Flags

- enhance and improve data
- enhance discoverability
- visualization aid
- transparency builds community trust
- facilitate potential development of automated updates by provider

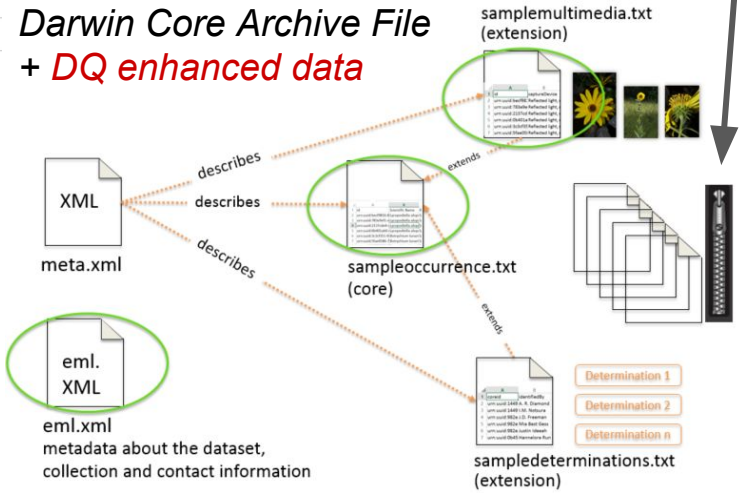


Data assurance practices

Flag	Records With This Flag	(%) Percent With This Flag
idigbio_isocountrycode_added	83287	99.74
dwc_continent_added	83258	99.705
dwc_phylum_added	82479	98.773
dwc_class_added	82471	98.763
dwc_order_added	82465	98.756
geopoint_datum_error	8223	9.847
geopoint_datum_missing	5137	6.152
rev_geocode_eez	691	0.828
rev_geocode_mismatch	249	0.298
rev_geocode_corrected		
rev_geocode_lon_sign		
geopoint_low_precision		



Darwin Core Archive File
+ DQ enhanced data



- Data Quality Flags *by recordset* at iDigBio
- Download **Darwin Core Archive Files**
  - *raw data*
  - *and a bonus file*
  - *available to everyone*
- **Annotations coming**
  - **feedback loop**
  - **transparency**

Contact the provider...

Specimen Record

Animalia > Chordata > Amphibia > Anura > Hyperolidae

Afrixalus fulvovittatus

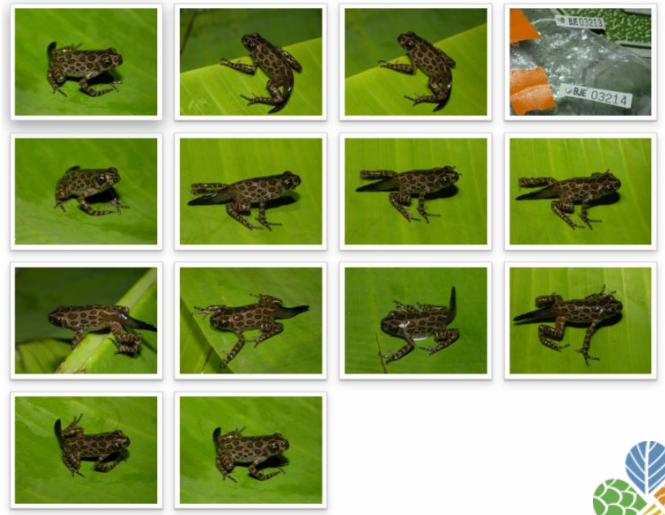
From Museum of Comparative Zoology, Harvard University

Continent	Africa	Institution Code	MCZ
Country	Cameroon	Collection Code	Herp
Locality	Nkambe	Catalog Number	A 148086
Latitude	6.5486166667		
Longitude	10.7599833333		

Contacts

Name	MCZ Harvard University	**Name**	Brendan Haley
Role	*none*	**Role**	Senior Database Manager
Email	*none*	**Email**	bhaley@oeb.harvard.edu
Phone	*none*	**Phone**	*none*

Media



Contents

- Summary
- Map
- Media
- Attribution
- All Data

https://www.idigbio.org/portal/records/d0199563-4490-4364-a768-e2318c3b4e24

Future DQ Work at iDigBio

Integrated Digitized Biocollections

- Full Taxonomic resolution against the GBIF Backbone, including: adding TaxonID (for all taxonomic levels) values, accepted names, and canonical names.
- Expanded geographic name/point validation/augmentation using GADM, GeoNames, or some similar place name database.
- Opening up the stage 1 corrections process to include externally provided corrections data sources. An API has been provisionally designed, but it hasn't undergone and testing or validation.
- Opening up the stage 1 corrections process to include annotations Since, from a technical perspective the process of applying an annotation to a record is the same as applying a general correction that only applies to specific single records or versions. This is even farther out than the last bullet, since it would need an entire workflow built around it to be effective.
- Harmonization of flag values, conditions and actions with other projects (such as this ALA/GBIF list: http://bit.ly/evMJv5)

NSF ADBC Digitization Thematic Collection Network

Plants, Herbivores, and Parasitoids: A Model System for the study of Tri-Trophic Associations

Photo: S. Bauer

Illustration: W. J. Hooker

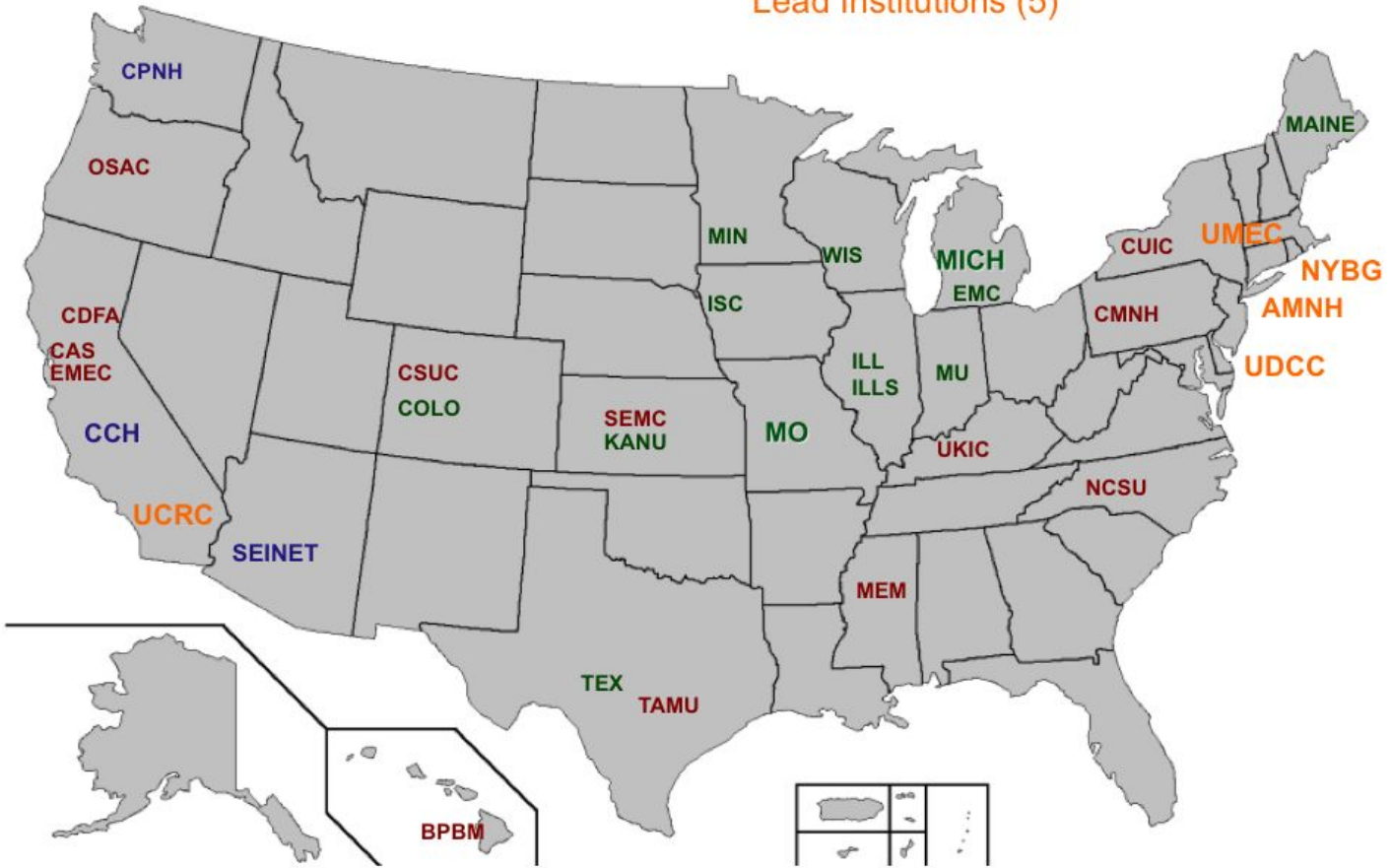
Photo: C. A. Johnson

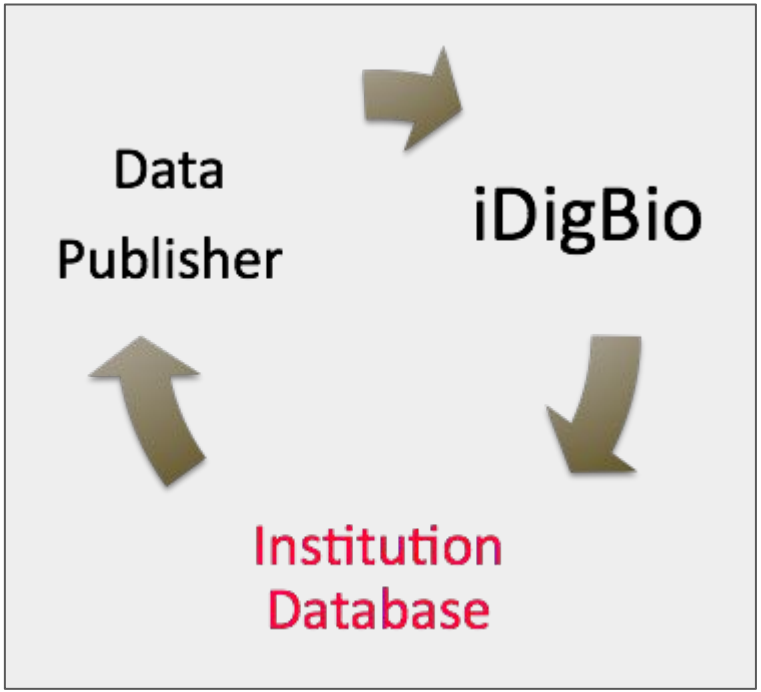
3,645,975

new records in 4 years

TTD Network

Herbaria (14) Botanical Data Contributors (3)
Insect Collections (18)
Lead Institutions (5)





1. Collection manager commitment and training
2. QC tools (data entry, data evaluation)
3. Research use of data

Acknowledgements and References

DRAFT FOR CONSULTATION: Best Practice Guide for Data Gap Analysis for biodiversity stakeholders http://www.gbif.org/system/files_force/gbif_resource/resource-82566/DRAFT-Best-Practice-Guide-for-Data-Gap-Analysis-for-biodiversity-stakeholders.pdf Arturo H. Ariño, Vishwas Chavan, Javier Otegui

VertNet Migrators https://github.com/VertNet/toolkit

VertNet Portal Spatial Quality Tab http://www.vertnet.org/resources/spatialqualitytabguide.html

VertNet GitHub Reference & Set Up for Data Issue Tracking http://www.vertnet.org/resources/issuetrackingguide.html

Improving Data Quality: iDigBio Recordset data cleaning method, tools, and data flags https://www.idigbio.org/content/improving-data-quality-idigbio-recordset-data-cleaning-method-tools-and-data-flags

A summer learning R to clean up data with the iDigBio portal recordset correction feature https://www.idigbio.org/content/summer-learning-r-clean-data-idigbio-portal-recordset-correction-feature

iDigBio Data recommendations for optimal searchability and applicability in the aggregate https://www.idigbio.org/wiki/index.php/Data_Ingestion_Guidance#Data_recommendations_for_optimal_searchability_and_applicability_in_the_aggregate

Exploring unique values in iDigBio using Apache Spark https://www.idigbio.org/content/exploring-unique-values-idigbio-using-apache-spark

Acknowledgements and References, cont'd

Biocode Field Information Management System ppt youtube. A Field Information Management System (FIMS) enables data collection at the source (in the field) by generating spreadsheet templates, validating data, and assigning persistent identifiers for every unique biological sample. The following diagram shows how the system works. The most typical functions are Generating Templates and Validating Data, both of which can be found under the Tools menu.

Generate a Template
Validate data
How FIMS works

Contact us

Deb Paul dpaul@fsu.edu
Katja Seltmann seltmann@ccber.ucsb.edu
Laura Russell larussell@vertnet.org
David Bloom dbloom@vertnet.org

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.