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

Deborah Paul, Katja Seltmann, Laura Russell, David Bloom

@idbdeb @iDigBio @VertNetOrg
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; ?, ...

**Possibilities**
- species ranges, outlier discovery, new species, gaps in collecting, relationships, predictive niche models, collector maps
Data, data types, data standards

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

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

[http://rs.tdwg.org/dwc/terms/index.htm](http://rs.tdwg.org/dwc/terms/index.htm)

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

DataONE Data Life Cycle

Plan

Collect

Assure

Describe

Preserve

Discover

Integrate

Analyze
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 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
General toolkit for working with VertNet data. We call these data "migrants." Once customized to an original data source, it converts original data into Darwin Core ready for upload to an Integrated Publishing Toolkit (IPT) resource. — Edit

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

<table>
<thead>
<tr>
<th>Row</th>
<th>class</th>
</tr>
</thead>
<tbody>
<tr>
<td>155</td>
<td>Insecta</td>
</tr>
<tr>
<td>156</td>
<td>LEPOSPONDYL</td>
</tr>
<tr>
<td>157</td>
<td>LISSAMPHIBIAN</td>
</tr>
<tr>
<td>158</td>
<td>Leptocardi</td>
</tr>
<tr>
<td>159</td>
<td>Lillopsida</td>
</tr>
<tr>
<td>160</td>
<td>Lycopodiopsida</td>
</tr>
<tr>
<td>161</td>
<td>MAMM INSECT TRACE</td>
</tr>
<tr>
<td>162</td>
<td>MAMM INVERT</td>
</tr>
<tr>
<td>163</td>
<td>MAMM PLANT</td>
</tr>
<tr>
<td>164</td>
<td>MAMM TRACE</td>
</tr>
<tr>
<td>165</td>
<td>MAMMALIA</td>
</tr>
</tbody>
</table>
## Data completeness

<table>
<thead>
<tr>
<th>Question</th>
<th>Status</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Are coordinates present?</td>
<td>✔</td>
<td>Yes</td>
</tr>
<tr>
<td>Is the country value present?</td>
<td>✔</td>
<td>Yes</td>
</tr>
<tr>
<td>Are both coordinates 0 (zero)?</td>
<td>✔</td>
<td>No</td>
</tr>
<tr>
<td>Do coordinates have three or more decimal figures?</td>
<td>✔</td>
<td>Yes</td>
</tr>
<tr>
<td>Do coordinates have datum?</td>
<td>✔</td>
<td>Yes</td>
</tr>
</tbody>
</table>

## Data inconsistencies

<table>
<thead>
<tr>
<th>Question</th>
<th>Status</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Are coordinates within specified country?</td>
<td>✔</td>
<td>Yes</td>
</tr>
<tr>
<td>Distance outside of specified country (in degrees)</td>
<td>✔</td>
<td>0</td>
</tr>
<tr>
<td>Distance outside of species range map (in degrees)</td>
<td>✔</td>
<td>Could not be assessed</td>
</tr>
</tbody>
</table>

## Data Errors

<table>
<thead>
<tr>
<th>Question</th>
<th>Status</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Is latitude between 90 and -90?</td>
<td>✔</td>
<td>Yes</td>
</tr>
<tr>
<td>Is longitude between 180 and -180?</td>
<td>✔</td>
<td>Yes</td>
</tr>
<tr>
<td>Are coordinates transposed?</td>
<td>✔</td>
<td>No</td>
</tr>
<tr>
<td>Is latitude hemisphere correct?</td>
<td>✔</td>
<td>Yes</td>
</tr>
<tr>
<td>Is longitude hemisphere correct?</td>
<td>✔</td>
<td>Yes</td>
</tr>
</tbody>
</table>

1Assessed 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!


Graphic by Joanna McCaffrey, iDigBio Biodiversity Informatics Manager
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

https://www.idigbio.org/portal/search
Data assurance practices

- 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 > Hyperoliidae

Africalus fulvovittatus
From Museum of Comparative Zoology, Harvard University

<table>
<thead>
<tr>
<th>Continent</th>
<th>Africa</th>
</tr>
</thead>
<tbody>
<tr>
<td>Country</td>
<td>Cameroon</td>
</tr>
<tr>
<td>Locality</td>
<td>Nkambe</td>
</tr>
<tr>
<td>Latitude</td>
<td>6.5486166667</td>
</tr>
<tr>
<td>Longitude</td>
<td>10.7598333333</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Institution</th>
<th>MCZ</th>
</tr>
</thead>
<tbody>
<tr>
<td>Code</td>
<td></td>
</tr>
<tr>
<td>Collection</td>
<td>Herp</td>
</tr>
<tr>
<td>Code</td>
<td></td>
</tr>
<tr>
<td>Catalog</td>
<td>A 148086</td>
</tr>
<tr>
<td>Number</td>
<td></td>
</tr>
</tbody>
</table>

Contacts

<table>
<thead>
<tr>
<th>Name</th>
<th>MCZ Harvard University</th>
</tr>
</thead>
<tbody>
<tr>
<td>Role</td>
<td>none</td>
</tr>
<tr>
<td>Email</td>
<td>none</td>
</tr>
<tr>
<td>Phone</td>
<td>none</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Name</th>
<th>Brendan Haley</th>
</tr>
</thead>
<tbody>
<tr>
<td>Role</td>
<td>Senior Database Manager</td>
</tr>
<tr>
<td>Email</td>
<td><a href="mailto:bhailey@oeb.harvard.edu">bhailey@oeb.harvard.edu</a></td>
</tr>
</tbody>
</table>

https://www.idigbio.org/portal/records/d0199563-4490-4364-a768-e2318c3b4e24
Future DQ Work at

- 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](http://bit.ly/evMJv5))
NSF ADBC Digitization Thematic Collection Network

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

3,645,975 new records in 4 years
1. Collection manager commitment and training

2. QC tools (data entry, data evaluation)

3. Research use of data
Acknowledgements and References


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

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


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.