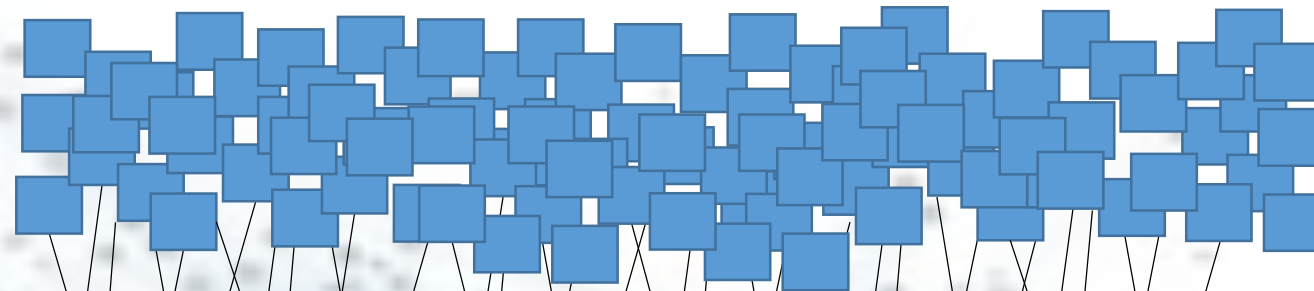


A faint, light blue background image showing a complex network of interconnected nodes and lines, resembling a molecular structure or a data network, centered behind the title text.

Practical use of aggregator data quality metrics in a collection scenario

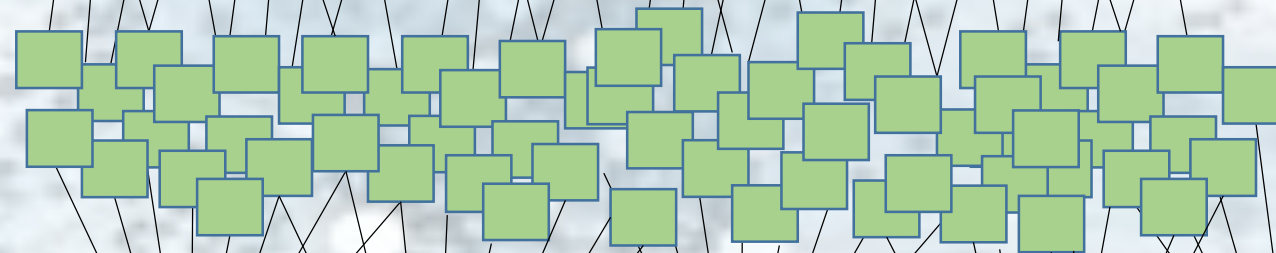
**Andrew Bentley
University of Kansas
Biodiversity Institute**

**Collections
Providers**



Darwin Core or ABCD

IPT servers



Aggregators



Publishing data to aggregators

Benefits:

- Visibility for collection and data
- Exposure to larger research and end user community
- Attribution for data usage
- Comparison with other collections

Leads to collections advocacy and increased use of collections and data


Data Quality

- **TDWG Biodiversity Data Quality Interest Group**
 - Standard set of Data Quality Tests and Assertions
- Assertions about data quality based on backbone taxonomic and geographic authorities and Darwin Core requirements
- Can be used by providers to check data quality, errors etc.
- Three aggregators providing data quality metrics
 - GBIF, iDigBio and ALA – Vertnet (GitHub)
- No standardization across aggregators as yet

Global Biodiversity Information Facility (GBIF)

KUBI Ichthyology Collection

OCCURRENCES PER REMARKS

Remarks	Count	
Coordinate rounded	9,976	
Identified date unlikely	4,584	
Country coordinate mismatch	509	
Taxon match higherrank	461	
Taxon match fuzzy	259	
Country derived from coordinates	170	
Geodetic datum assumed WGS84	37	
Coordinate uncertainty meters invalid	4	
Taxon match none	4	
Country invalid	2	

NEXT

Taxon match fuzzy	259	
Country derived from coordinates	170	
Geodetic datum assumed WGS84	37	
Coordinate uncertainty meters invalid	4	
Taxon match none	4	
Country invalid	2	

NEXT



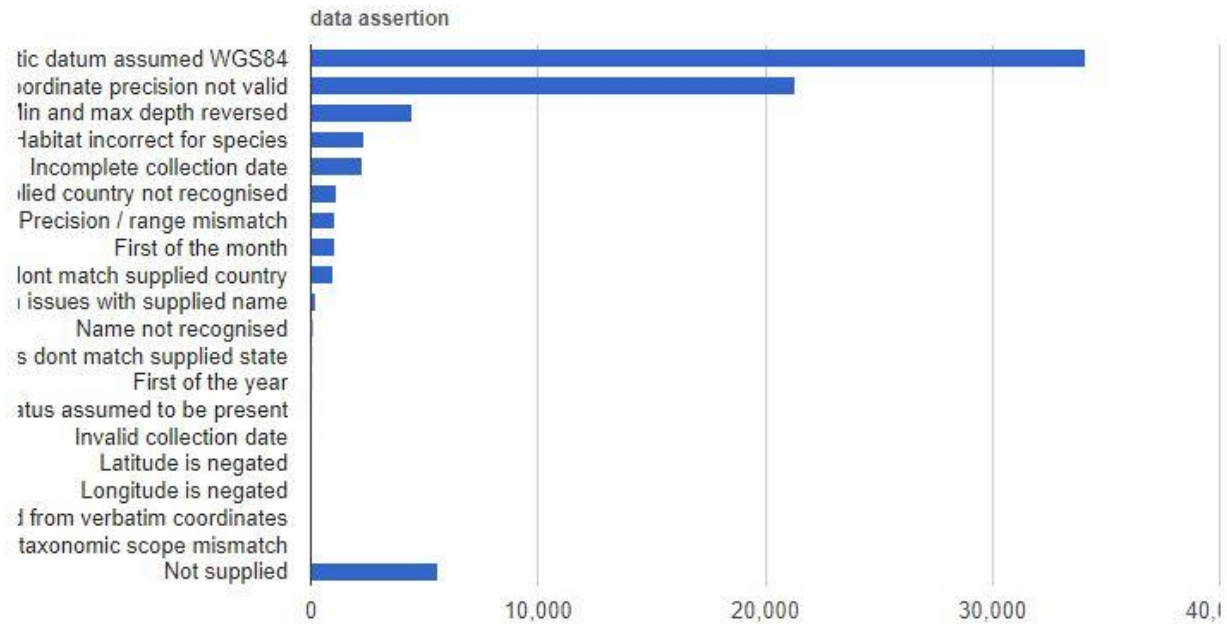
745 other or unknown

Atlas of Living Australia (ALA)



Search the Atlas ...

Start exploring ▾ Search & analyse ▾ Participate ▾ Learn about the ALA ▾



Integrated Digitized Biocollections (iDigBio)

Data Corrected Data Use Raw

This table shows any data corrections that were performed on this recordset to improve the capabilities of iDigBio Search. The first column represents the correction performed. The last two columns represent the number and percentage of records that were corrected. A complete list of the data quality flags and their descriptions can be found [here](#). Clicking on a data flag name will take you to a search for all records with this flag in this recordset.

Flag	Records With This Flag	(%) Percent With This Flag
dwc_basISOofrecord_invalid 1	10982	100
dwc_basISOofrecord_removed 1	10982	100
idigbio_isocountrycode_added 1	10528	95.888
dwc_datasetid_added 1	10401	94.71
dwc_parentnameusageid_added 1	10401	94.71
dwc_taxonid_added 1	10401	94.71
dwc_taxonomicstatus_added 1	10401	94.71
dwc_taxonrank_added 1	10401	94.71
gbif_canonicalname_added 1	10401	94.71
gbif_genericname_added 1	10401	94.71
gbif_taxon_corrected 1	10401	94.71
gbif_reference_added 1	10332	94.081
gbif_vernacularname_added 1	10104	92.006
dwc_multimedia_added 1	7241	65.936
dwc_scientificnameauthorship_replaced 1	6872	62.576
dwc_originalnameusageid_added 1	5755	52.404
rev_geocode_eez 1	5742	52.288
dwc_continent_replaced 1	3185	29.002
dwc_scientificnameauthorship_added 1	2120	19.304
taxon_match_failed 1	1125	10.244
rev_geocode_mismatch 1	989	8.824
dwc_family_replaced 1	442	4.025
dwc_order_replaced 1	404	3.679
dwc_specificepithet_replaced 1	279	2.541
dwc_genus_replaced 1	221	2.012
geopoint_low_precision 1	166	1.421
dwc_taxonremarks_added 1	66	0.601
rev_geocode_failure 1	37	0.337
dwc_class_replaced 1	35	0.319
rev_geocode_corrected 1	24	0.219
rev_geocode_lon_sign 1	24	0.219
rev_geocode_eez_corrected 1	10	0.091
dwc_genus_added 1	8	0.073
dwc_specificepithet_added 1	5	0.046
dwc_phylum_replaced 1	3	0.027
geopoint_bounds 1	1	0.009

Is this normal? Comparison to other collections

Data Corrected	Data Use	Raw
<p>This table shows any data corrections that were performed on this recordset to improve the capabilities of iDigBio Search. The first column represents the correction performed. The last two columns represent the number and percentage of records that were corrected. A complete list of the data quality flags and their descriptions can be found here. Clicking on a data flag name will take you to a search for all records with this flag in this recordset.</p>		
Flag	Records With This Flag	(%) Percent With This Flag
idigbio_isocountrycode_added	207768	98.558
dwc_datASETid_added	204559	97.038
dwc_parentnameusageid_added	204559	97.038
dwc_taxonid_added	204559	97.038
dwc_taxonomicstatus_added	204559	97.038
gbif_canonicalname_added	204559	97.038
gbif_genericname_added	204559	97.038
gbif_taxon_corrected	204559	97.038
dwc_scientificnameauthorship_added	201185	95.438
gbif_vernacularname_added	200831	95.173
gbif_reference_added	197940	93.895
dwc_multimedia_added	189131	89.718
dwc_taxonrank_replaced	153430	72.782
dwc_originalnameusageid_added	93543	44.374
geopoint_datum_error	75408	35.771
geopoint_low_precision	38102	17.128
taxon_match_failed	15702	7.449
rev_geocode_eez	9946	4.718
dwc_family_replaced	8428	3.997
dwc_genus_replaced	8751	3.202
dwc_specificepithet_replaced	5785	2.735
dwc_infraspecificepithet_added	3643	1.87
rev_geocode_mismatch	2850	1.257
dwc_continent_replaced	1388	0.683
dwc_taxonremarks_added	793	0.378
rev_geocode_corrected	531	0.252
rev_geocode_lon_sign	432	0.205
dwc_kingdom_suspect	384	0.173
dwc_country_replaced	298	0.141
dwc_stateprovince_replaced	297	0.141
dwc_infraspecificepithet_replaced	224	0.108
dwc_continent_added	188	0.089
rev_geocode_failure	137	0.065
rev_geocode_lat_sign	89	0.042
geopoint_similar_coord	57	0.027
geopoint_datum_missing	49	0.023
dwc_order_replaced	42	0.02
rev_geocode_eez_corrected	17	0.008
dwc_taxonremarks_replaced	14	0.007
rev_geocode_flip	8	0.003
dwc_specificepithet_added	4	0.002
rev_geocode_flip_both_sign	4	0.002
datecollected_bounds	2	0.001

This table shows any data corrections that were performed on this recordset to improve the capabilities of iDigBio Search. The first column represents the correction performed. The last two columns represent the number and percentage of records that were corrected. A complete list of the data quality flags and their descriptions can be found [here](#). Clicking on a data flag name will take you to a search for all records with this flag in this recordset.

Flag	Records With This Flag	(%) Percent With This Flag
dwc_basISOrecord_invalid 1	10982	100
dwc_basISOrecord_removed 1	10982	100
idigbio_isocountrycode_added 1	10528	95.866
dwc_datasetid_added 1	10408	94.773
dwc_parentnameusageid_added 1	10408	94.773
dwc_taxonid_added 1	10408	94.773
dwc_taxonomicstatus_added 1	10408	94.773
dwc_taxonrank_added 1	10408	94.773
gbif_canonicalname_added 1	10408	94.773
gbif_genericname_added 1	10408	94.773
gbif_taxon_corrected 1	10408	94.773
gbif_reference_added 1	10339	94.145
gbif_vernacularname_added 1	10110	92.06
dwc_multimedia_added 1	7248	65.999
dwc_scientificnameauthorship_replaced 1	6877	62.821
dwc_country_replaced 1	6009	54.717
dwc_originalnameusageid_added 1	5753	52.386
rev_geocode_eez 1	5742	52.286
dwc_continent_replaced 1	3185	29.002
dwc_scientificnameauthorship_added 1	2128	19.377
taxon_match_failed 1	1117	10.171
rev_geocode_mismatch 1	989	8.824
dwc_family_replaced 1	450	4.098
dwc_order_replaced 1	413	3.761
dwc_specificepithet_replaced 1	342	3.114
dwc_genus_replaced 1	229	2.085
geopoint_low_precision 1	156	1.421
dwc_taxonremarks_added 1	86	0.801
rev_geocode_failure 1	37	0.337
dwc_class_replaced 1	35	0.319
rev_geocode_corrected 1	24	0.219
rev_geocode_lon_sign 1	24	0.219
rev_geocode_eez_corrected 1	10	0.091
dwc_genus_added 1	8	0.073
dwc_specificepithet_added 1	5	0.046
geopoint_datum_missing 1	5	0.046
dwc_phylum_replaced 1	3	0.027
geopoint_bounds 1	1	0.009

Collection

Taxonomy

Geography

Multimedia

non

Logical categories


- **Metrics I can do “nothing” about**
 - Aggregator specific augmenting fields
 - Fields I am unable or unwilling to map to in Darwin Core given my CMS
- **Metrics I can do something about**
 - Things I can attend to short term – “easy”
 - Things I can attend to long term – “hard”
- **Differences of opinion or errors**
 - Taxonomic and geographic authority anomalies or conflicts

Metrics I can do “nothing” about

- idigbio_isocountrycode_added
- dwc_datasetid_added
- dwc_parentnameusageid_added
- dwc_taxonid_added
- dwc_taxonomicstatus_added
- gbif_canonicalname_added
- gbif_genericname_added
- gbif_reference_added
- gbif_vernacularname_added
- dwc_originalnameusageid_added


Metrics I can do something about


- 1
- 2
- 3
- 4
- 5
- 6
- 7
- 8
- 9
- 10
- 11
- 12
- 13
- 14
- 15
- 16
- 17
- 18
- 19
- 20
- 21
- 22
- 23
- 24
- 25
- 26
- 27
- 28
- 29
- 30
- 31
- 32
- 33
- 34
- 35
- 36
- 37
- 38
- 39
- 40
- 41
- 42
- 43
- 44
- 45
- 46
- 47
- 48
- 49
- 50
- 51
- 52
- 53
- 54
- 55
- 56
- 57
- 58
- 59
- 60
- 61
- 62
- 63
- 64
- 65
- 66
- 67
- 68
- 69
- 70
- 71
- 72
- 73
- 74
- 75
- 76
- 77
- 78
- 79
- 80
- 81
- 82
- 83
- 84
- 85
- 86
- 87
- 88
- 89
- 90
- 91
- 92
- 93
- 94
- 95
- 96
- 97
- 98
- 99
- 100





abentley | KU Fish Teaching Collection ▾


Notifications: 0


 Data Entry


 Interactions


 Trees

 Record Sets

 Queries

 Reports

 Attachments

 WorkBench

App Resource

Global Resources

Disciplines

Ichthyology

Borrow Invoice Report

Datamax Jar Labels

Datamax Jar Labels

Datamax Jar Labels

DataObjFormatters

Fish Gift Report

Fish Loan Report

Gift Shipping Form

Loan Shipping Form

Teaching labels

Teaching labels - no family

Tissue Gift Report

UIFormatters

WebLinks

New Resource

KU Fish Voucher Collection

KU Fish Tissue Collection

DwCA_tissue

DwCA_tissue_metadata

Number Tags

WB Tag Label

New Resource

USER TYPES

USERS

KU Fish Teaching Collection

```
<field term="http://rs.tdwg.org/dwc/terms/continent" isNot="false" isRelFld="false" oper="11" stringId="1,10,2,3.geography.Continent" value=""/>
<field term="http://rs.tdwg.org/dwc/terms/country" isNot="false" isRelFld="false" oper="11" stringId="1,10,2,3.geography.Country" value=""/>
<field term="http://rs.tdwg.org/dwc/terms/stateProvince" isNot="false" isRelFld="false" oper="11" stringId="1,10,2,3.geography.State" value=""/>
<field term="http://rs.tdwg.org/dwc/terms/county" isNot="false" isRelFld="false" oper="11" stringId="1,10,2,3.geography.County" value=""/>
<field term="http://rs.tdwg.org/dwc/terms/higherGeography" isNot="false" isRelFld="true" oper="11" stringId="1,10,2,3.geography.geography" value=""/>

<!-- Locality Detail -->
<field term="http://rs.tdwg.org/dwc/terms/waterBody" isNot="false" isRelFld="false" oper="11" stringId="1,10,2,124-localityDetails.localitydetail.waterBody" value=""/>
<field term="http://rs.tdwg.org/dwc/terms/island" isNot="false" isRelFld="false" oper="11" stringId="1,10,2,124-localityDetails.localitydetail.island" value=""/>
<field term="http://rs.tdwg.org/dwc/terms/islandGroup" isNot="false" isRelFld="false" oper="11" stringId="1,10,2,124-localityDetails.localitydetail.islandGroup" value=""/>
</query>
</queries>
</core>

<!-- GGBN Amplification Extension -->
<extension rowType="http://data.ggbn.org/schemas/ggbn/terms/Amplification">
  <queries>
    <query name="amplification.csv" contextTableId="1">
      <id term="http://rs.tdwg.org/dwc/terms/occurrenceID" isNot="false" isRelFld="false" stringId="1.collectionobject.guid" oper="11" value=""/>
      <field term="http://data.ggbn.org/schemas/ggbn/terms/BOLDProcessID" isNot="false" isRelFld="false" oper="11" stringId="1,121-dnaSequences.dnasequence.boldBarcodeId" value=""/>
      <field term="http://data.ggbn.org/schemas/ggbn/terms/geneticAccessionNumber" isNot="false" isRelFld="false" oper="11" stringId="1,121-dnaSequences.dnasequence.text2" value=""/>
      <field term="http://data.ggbn.org/schemas/ggbn/terms/geneticAccessionURI" isNot="false" isRelFld="false" oper="11" stringId="1,121-dnaSequences.dnasequence.text1" value=""/>
      <field term="http://data.ggbn.org/schemas/ggbn/terms/marker" isNot="true" isRelFld="false" oper="12" stringId="1,121-dnaSequences.dnasequence.moleculeType" value=""/>
      <field term="http://data.ggbn.org/schemas/ggbn/terms/amplificationStaff" isNot="false" isRelFld="true" oper="11" stringId="1,121-dnaSequences.s-sequencer.agent.sequencer" value=""/>
    </query>
  </queries>
</extension>

<!-- Audubon Core Extension -->
<extension rowType="http://rs.tdwg.org/ac/terms/Multimedia">
  <field term="http://purl.org/dc/terms/accessRights" value="http://biodiversity.ku.edu/research/university-kansas-biodiversity-institute-data-publication-and-use-norms/">
  <field term="http://ns.adobe.com/xap/1.0/rights/Owner" value="University of Kansas Biodiversity Institute"/>
  <field term="http://ns.adobe.com/photoshop/1.0/Credit" value="KU Biodiversity Institute Ichthyology Division"/>
  <field term="http://purl.org/dc/terms/rights" value="https://creativecommons.org/licenses/by/4.0/deed.en_US"/>
  <field term="http://rs.tdwg.org/ac/terms/licenseLogoURL" value="https://licensebuttons.net/1/by/4.0/88x31.png"/>
  </field>
</extension>

<queries>
  <!-- Collection Object Attachment -->
  <query name="COAudubonCore.csv" contextTableId="1">
    <id term="http://rs.tdwg.org/dwc/terms/occurrenceID" isNot="false" isRelFld="false" stringId="1.collectionobject.guid" oper="11" value=""/>
    <field term="http://purl.org/dc/terms/accessRights" value="http://biodiversity.ku.edu/research/university-kansas-biodiversity-institute-data-publication-and-use-norms/">
    <field term="http://ns.adobe.com/xap/1.0/rights/Owner" value="University of Kansas Biodiversity Institute"/>
    <field term="http://ns.adobe.com/photoshop/1.0/Credit" value="KU Biodiversity Institute Ichthyology Division"/>
    <field term="http://purl.org/dc/terms/rights" value="https://creativecommons.org/licenses/by/4.0/deed.en_US"/>
    <field term="http://rs.tdwg.org/ac/terms/licenseLogoURL" value="https://licensebuttons.net/1/by/4.0/88x31.png"/>
  </query>
</queries>

</core>
```

Save

Delete

Data Corrected Data Use Raw

This table shows any data corrections that were performed on this recordset to improve the capabilities of iDigBio Search. The first column represents the correction performed. The last two columns represent the number and percentage of records that were corrected. A complete list of the data quality flags and their descriptions can be found [here](#). Clicking on a data flag name will take you to a search for all records with this flag in this recordset.

Flag	Records With This Flag	(%) Percent With This Flag
idigbio_isocountrycode_added	41235	98.829
dwc_datasetid_added	41176	98.488
dwc_parentnameusageid_added	41176	98.488
dwc_taxonid_added	41176	98.488

OCCURRENCES PER REMARKS

Remarks	Count	
Coordinate rounded	9,976	<div></div>
Identified date unlikely	4,584	<div></div>
Country coordinate mismatch	509	<div></div>
Taxon match higherrank	461	<div></div>
Taxon match fuzzy	259	<div></div>
Country derived from coordinates	170	<div></div>
Geodetic datum assumed WGS84	37	<div></div>
Coordinate uncertainty meters invalid	4	<div></div>
Taxon match none	4	<div></div>
Country invalid	2	<div></div>

NEXT

OCCURRENCES PER ISSUES AND FLAGS

Issues and flags	Count	
Coordinate rounded	9,975	<div></div>
Identified date unlikely	4,584	<div></div>
Country coordinate mismatch	509	<div></div>
Taxon match higherrank	379	<div></div>
Country derived from coordinates	168	<div></div>
Taxon match fuzzy	7	<div></div>
Presumed negated latitude	1	
Presumed negated longitude	1	

dwc_country_replaced	8	0.019
dwc_infraspecificpithet_added	4	0.01
dwc_kingdom_suspect	4	0.01
dwc_specificpithet_added	3	0.007
rev_geocode_both_sign	1	0.002
rev_geocode_corrected	1	0.002
rev_geocode_eez_corrected	1	0.002

Specimen Record

Citation	Uploads	Related species
----------	---------	-----------------



[iDigBio Home](#)
[Portal Home](#)
[Search Records](#)
[Learning Center](#)
[Data](#)
[Research Collaboration](#)
[Feedback](#)

Psilorhynchidae	Psilorhynchus sucatio	2008-11-02	Nepal
Lepisosteidae	Lepisosteus platostomus	2007-05-17	United States
Centrarchidae	Lepomis cyanellus x L. macrochirus	2006-08-15	United States
Centrarchidae	Lepomis cyanellus x L. macrochirus	2006-08-07	United States
Centrarchidae	Lepomis cyanellus x L. macrochirus	2006-07-11	United States
Percidae	Etheostoma flabellare	2006-07-06	United States
Percidae	Etheostoma flabellare	2006-07-01	United States
Centrarchidae	Lepomis cyanellus x L. macrochirus	2006-06-19	United States

994:199 [ref. [21581](#)], warren et al. 1994:132 [ref. [25836](#)], Murdy et al. 1997:57 [ref. [23144](#)], ss et al. 2001:93 [ref. [25978](#)], Nelson et al. Scharpf 2005:8 [ref. [28940](#)], Page & Burr æe. Distribution: Eastern North America; introduced

General problems/improvements

iDigBio / idigbio-search-api
[Watch](#)

<> Code
Issues 13
Pull requests 0
Projects 0
Wiki
Insights

Data Quality Flags

Dan Stoner edited this page on Nov 30, 2017 · 30 revisions

Purpose

This document describes how iDigBio identifies known data quality issues of ingested specimen data and represents them in the iDigBio Search API. During the ingestion process, iDigBio often encounters data that are missing, factually incorrect, or out of compliance with meta-data standards and controlled vocabularies. For example, Taxonomic Names are added from the [GBIF Backbone Taxonomy](#). To facilitate indexing, corrections are made to these data and they are flagged in the search API.

Flags

The table below describes the flags currently used by iDigBio:

Flag	Definition
datecollected_bounds	Date Collected out of bounds (1700-01-02, Date of Indexing).
dwc_basisofrecord_paleo_conflict	Basis of Record was not FossilSpecimen, but the record contains paleo context terms.
dwc_class_added	Darwin Core Class Added. http://terms.tdwg.org/wiki/dwc:class
dwc_class_replaced	Darwin Core Class Corrected.
dwc_continent_added	Darwin Core Continent Added. http://terms.tdwg.org/wiki/dwc:continent
dwc_continent_replaced	Darwin Core Continent Corrected.
dwc_country_added	Darwin Core Country Added. http://terms.tdwg.org/wiki/dwc:country
dwc_country_replaced	Darwin Core Country Corrected.
dwc_kingdom_added	Darwin Core Kingdom Added. http://terms.tdwg.org/wiki/dwc:kingdom

Conclusions

- Some DQ metrics very useful
- Some DQ metrics not actionable
- Some DQ metrics confusing – resolved through better descriptions
- Take everything with a healthy dose of skepticism and use all available resources to assist in checking accuracy of metrics.
- Would be great if we could take the best of each aggregators approach and create standardized metrics and UI representation.

How to get data back into database

- **Correcting records one-by-one very time consuming.**
- **Specify Software in discussions with GBIF and others about possibility of incorporating data quality metric API into the database itself so that data quality checks could be performed before publishing and corrections made at the source in some semi-automated fashion.**

Acknowledgements

- GBIF
- iDigBio
- Deb Paul and Nicole Fisher

Thank you

