Who Has Time for Biological Collections Data Quality Feedback? Maybe a Community Can Help

Teresa J. Mayfield-Meyer
Collections Assistant, Museum of Southwestern Biology
Treasurer, Arctos Consortium

SPNHC/TDWG 2018

Website: http://arctosdb.org
Search: http://arctos.database.museum

What do biodiversity collection managers do all day?

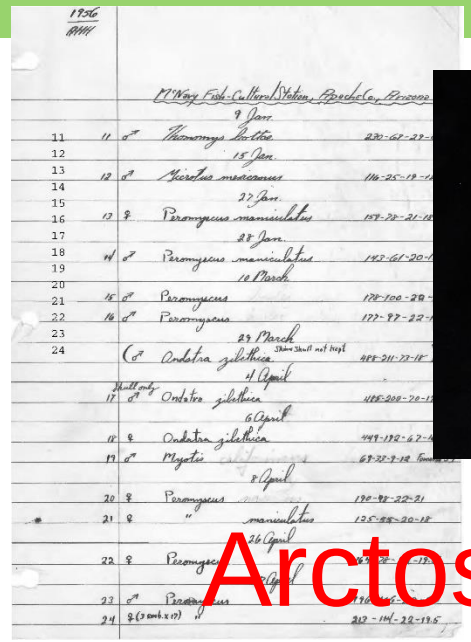
Arctos

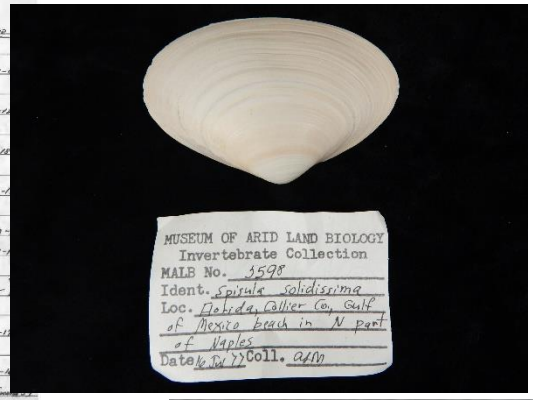
What do biodiversity collection managers do all day?

DIGITIZE AND MANAGE DATA

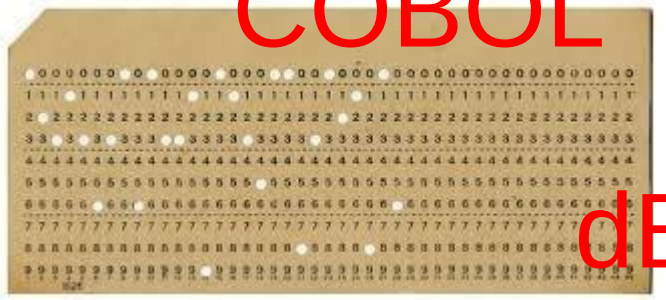
Arctos

Stuff happens when you digitize…





COBOL



dBASE

Arctos

EMu

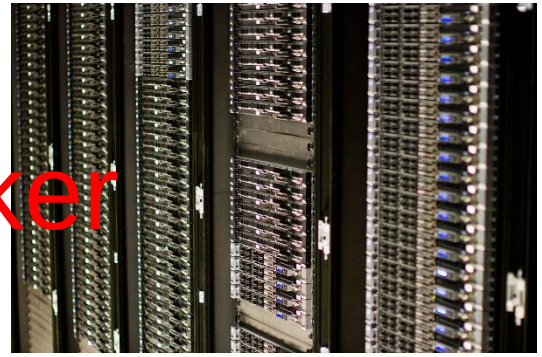
Specify

Symbiota

ACCESS

LOTUS123

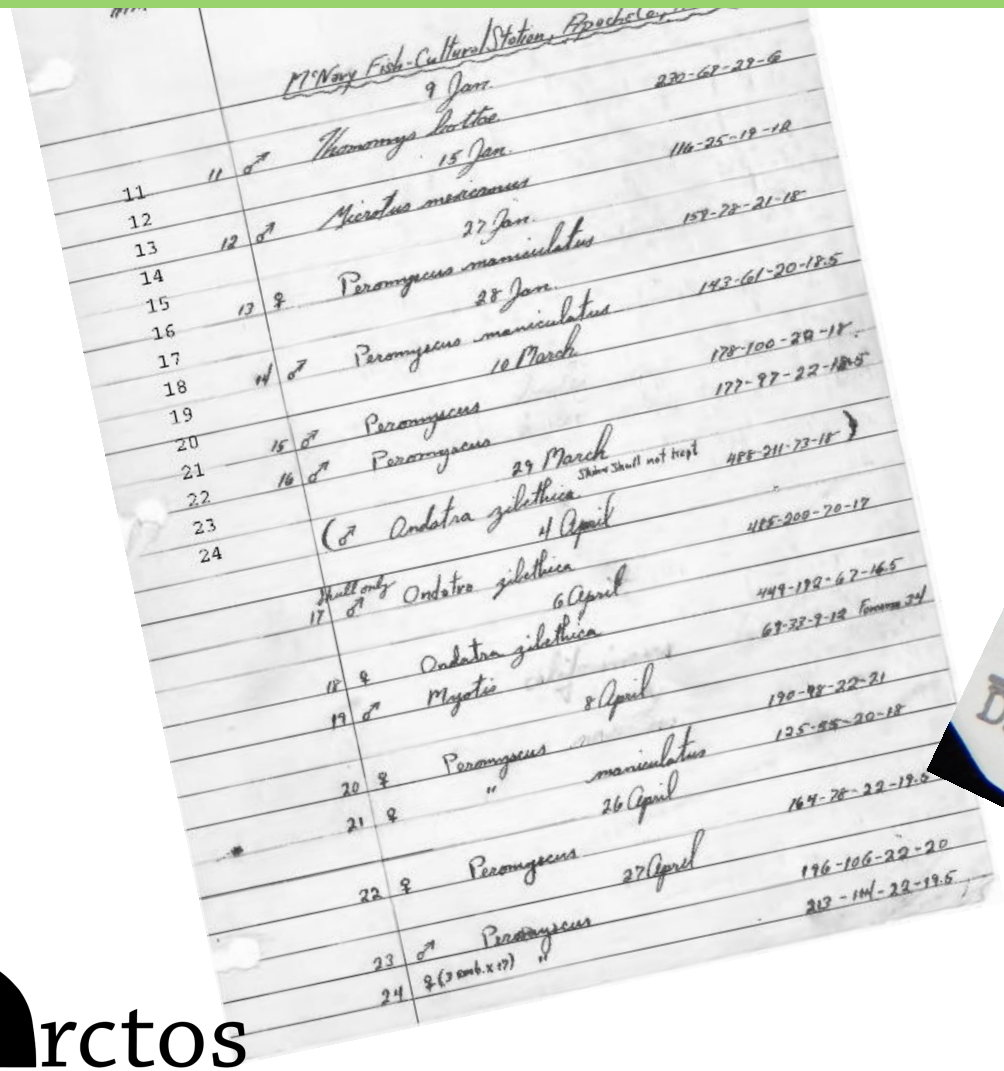
FileMaker

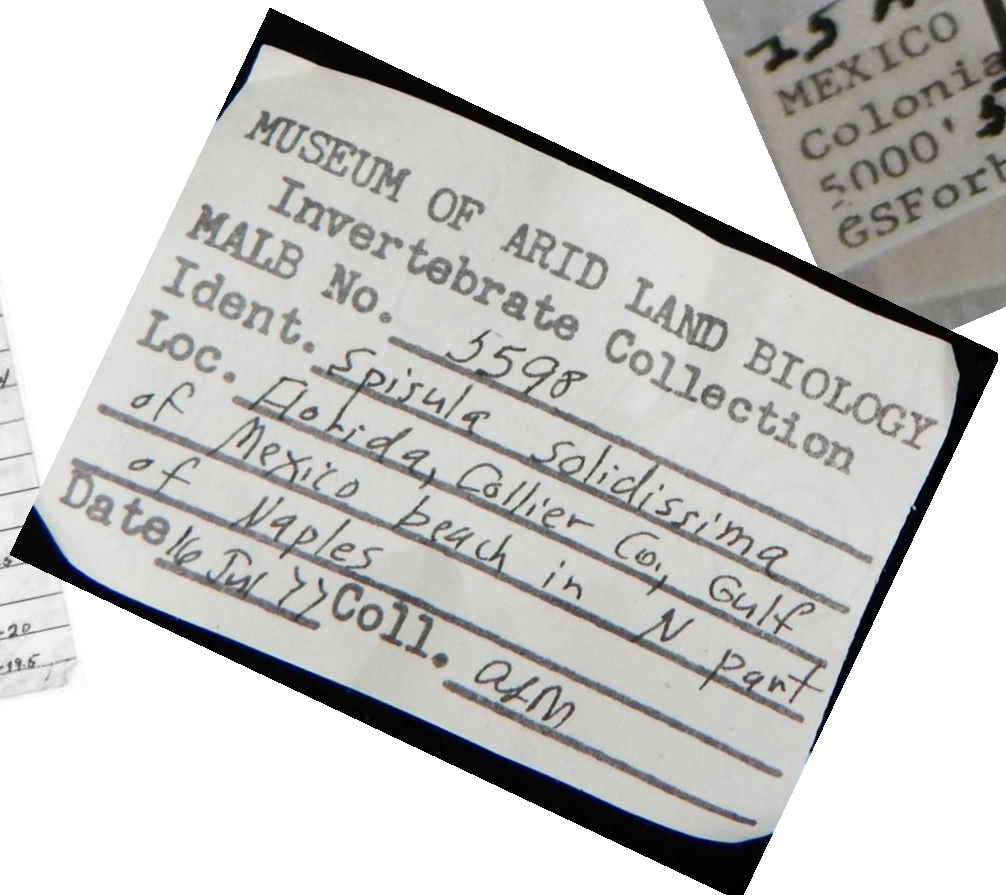


Cloud Computing

Arctos

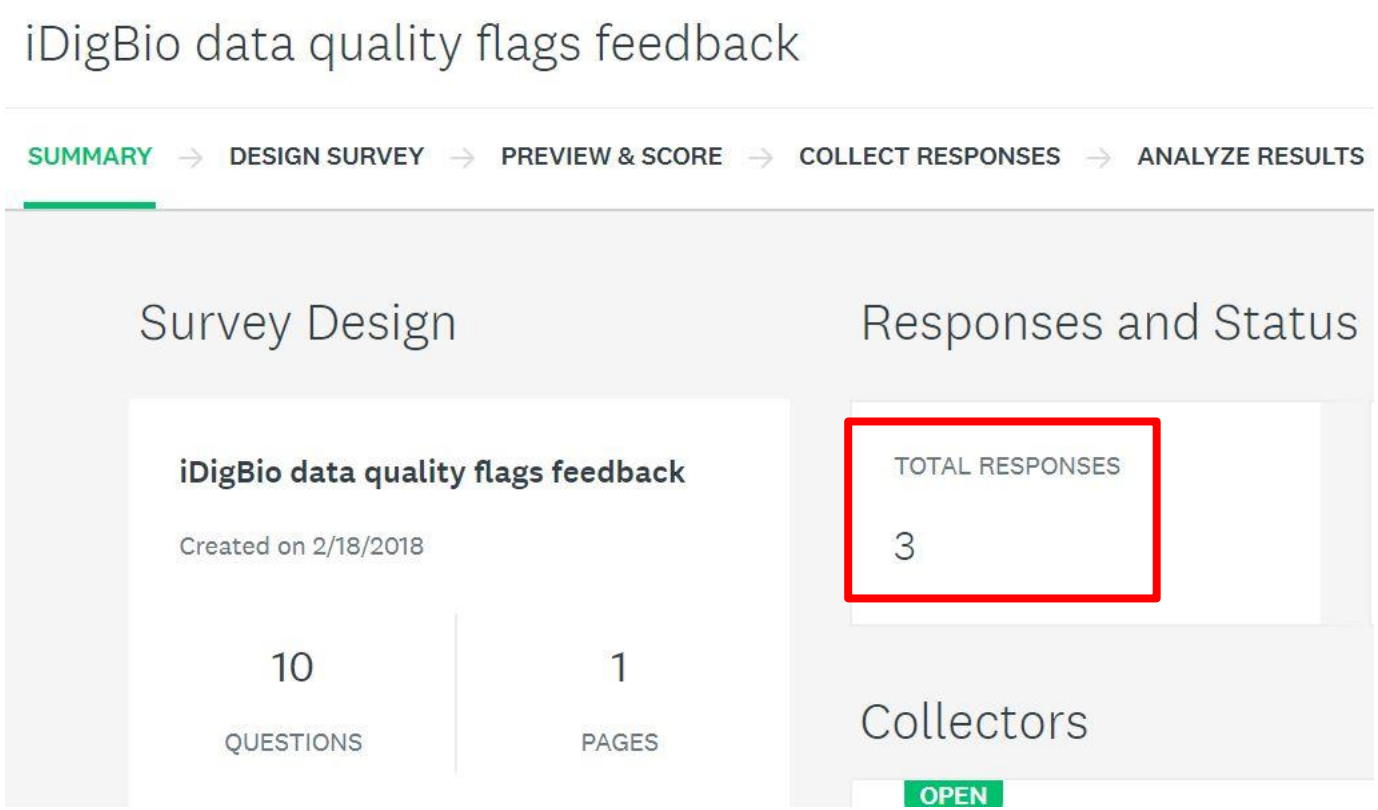
No Excuses (Not)





Arctos

Who has time for surveys?

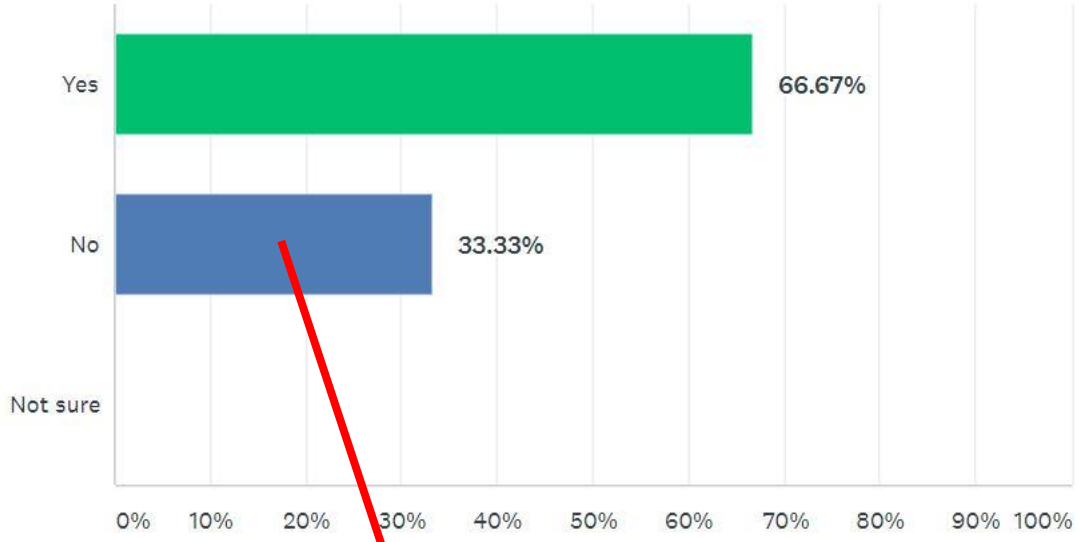


Arctos

Data Publishing

Have you published your collection data to iDigBio?

Answered: 3 Skipped: 0



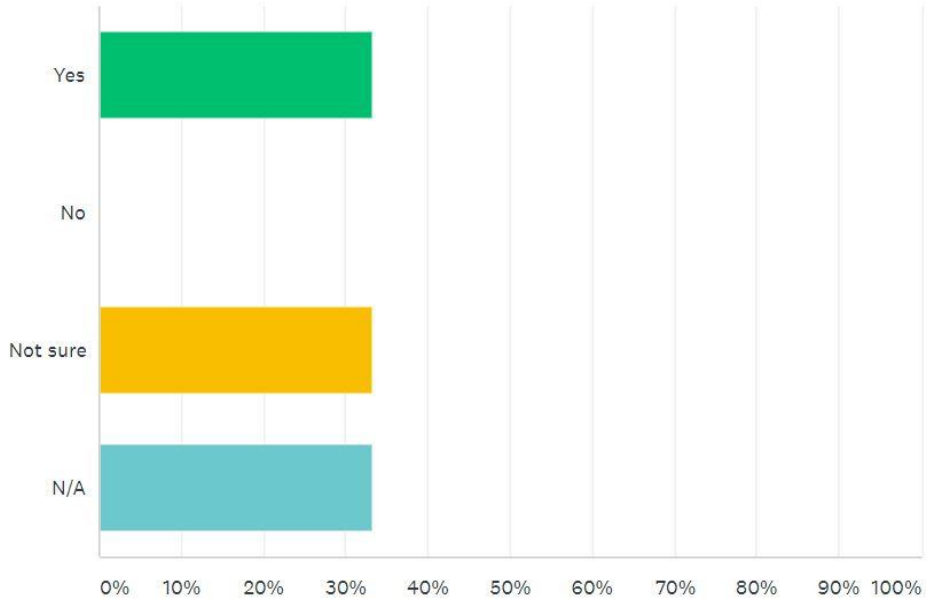
I'm not even sure what this is referencing:
data in Arctos or data that is being migrated into Arctos! I have never heard this terminology until now.

Arctos

Wait, there are data flags?

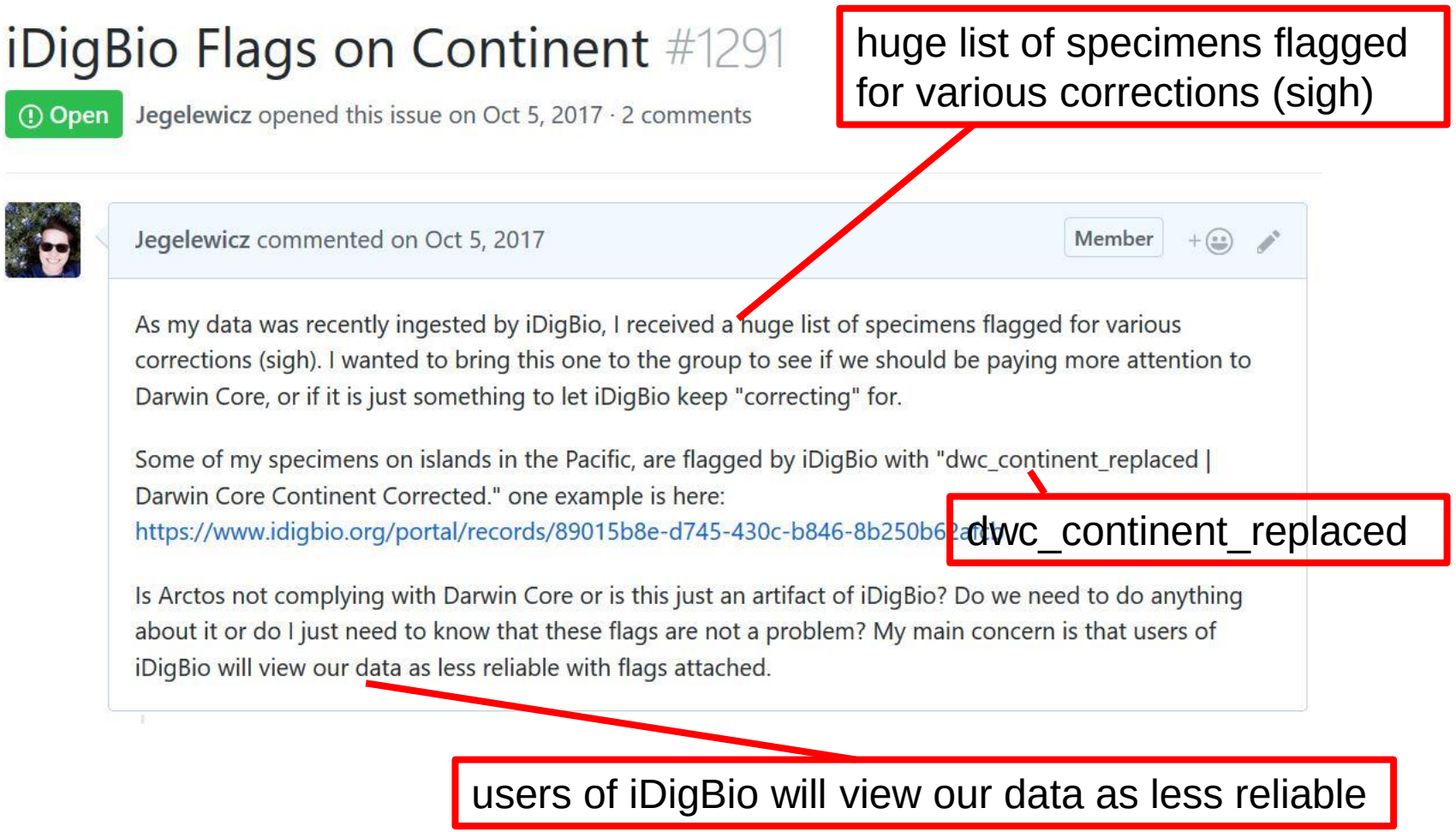
Whether or not you received a data quality email from iDigBio, have you reviewed the data quality flags provided by iDigBio for your collection data?

Answered: 3 Skipped: 0



Arctos

My initial experience



Arctos

My initial experience

noticed something similar in GBIF regarding dates

Darwin Core is an exchange standard; Arctos isn't "complying" with any data standards because none exist

iDigBio Flags on Continent #1291

Open Jegelewicz opened this issue on Oct 5, 2017 · 2 comments

Jegelewicz commented on Oct 5, 2017 Member

As my data was recently ingested by iDigBio, I received a huge list of specimens flagged for various corrections (sigh). I wanted to bring this one to the group to see if we should be paying more attention to Darwin Core, or if it is just something to let iDigBio keep "correcting" for.

Some of my specimens on islands in the Pacific, are flagged by iDigBio with "dwc_continent_replaced | Darwin Core Continent Corrected." one example is here:
https://www.idigbio.org/portal/records/89015b8e-d745-430c-b846-8b250b62afcb

Is Arctos not complying with Darwin Core or is this just an artifact of iDigBio? Do we need to do anything about it or do I just need to know that these flags are not a problem? My main concern is that users of iDigBio will view our data as less reliable with flags attached.

I agree with your assessment: User's initial reaction to the flag will be "Arctos is broken," which is absolutely not the case.

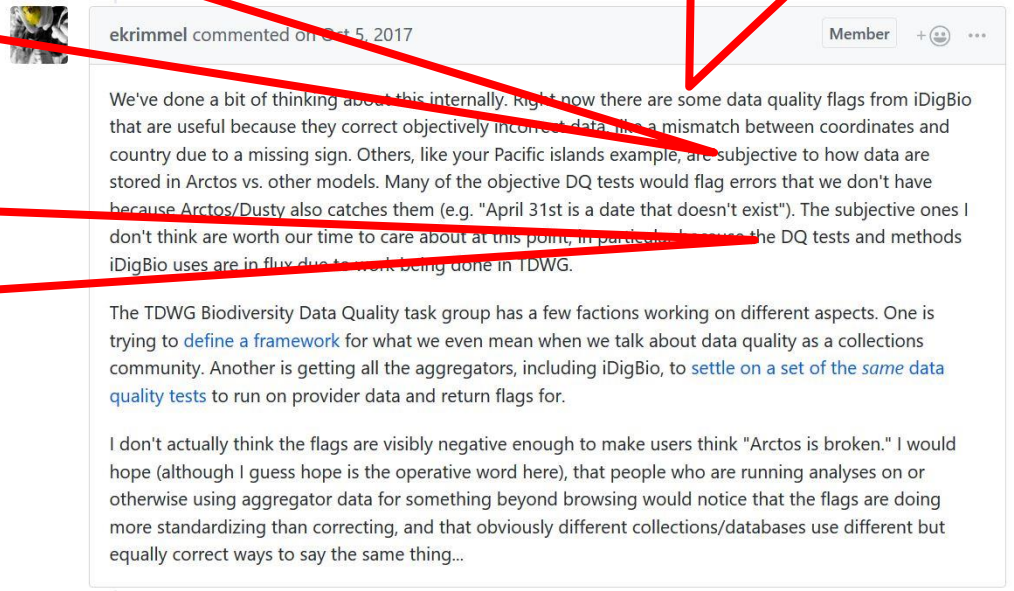
Arctos

My initial experience

"..subjective to how data are stored in Arctos vs. other models. The subjective ones I don't think are worth our time to care about at this point"

"..some data quality flags from iDigBio that are useful because they correct objectively incorrect data"

the DQ tests and methods iDigBio uses are in flux due to work being done in TDWG. ….



ekrimmel commented on Oct 5, 2017 — Member

We've done a bit of thinking about this internally. Right now there are some data quality flags from iDigBio that are useful because they correct objectively incorrect data, like a mismatch between coordinates and country due to a missing sign. Others, like your Pacific islands example, are subjective to how data are stored in Arctos vs. other models. Many of the objective DQ tests would flag errors that we don't have because Arctos/Dusty also catches them (e.g. "April 31st is a date that doesn't exist"). The subjective ones I don't think are worth our time to care about at this point, in particular because the DQ tests and methods iDigBio uses are in flux due to work being done in TDWG.

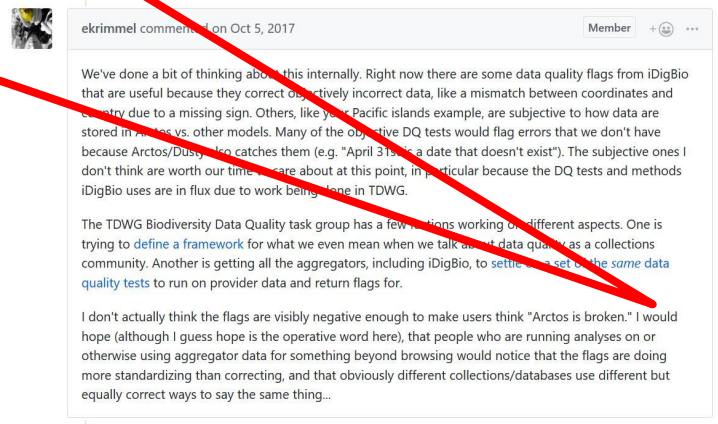
The TDWG Biodiversity Data Quality task group has a few factions working on different aspects. One is trying to define a framework for what we even mean when we talk about data quality as a collections community. Another is getting all the aggregators, including iDigBio, to settle on a set of the *same* data quality tests to run on provider data and return flags for.

I don't actually think the flags are visibly negative enough to make users think "Arctos is broken." I would hope (although I guess hope is the operative word here), that people who are running analyses on or otherwise using aggregator data for something beyond browsing would notice that the flags are doing more standardizing than correcting, and that obviously different collections/databases use different but equally correct ways to say the same thing...

Arctos

My initial experience

I would hope (although I guess hope is the operative word here), **that people who are running analyses on or otherwise using aggregator data for something beyond browsing would notice that the flags are doing more standardizing than correcting, and that obviously different collections/databases use different but equally correct ways to say the same thing...**



ekrimmel commented on Oct 5, 2017 — Member

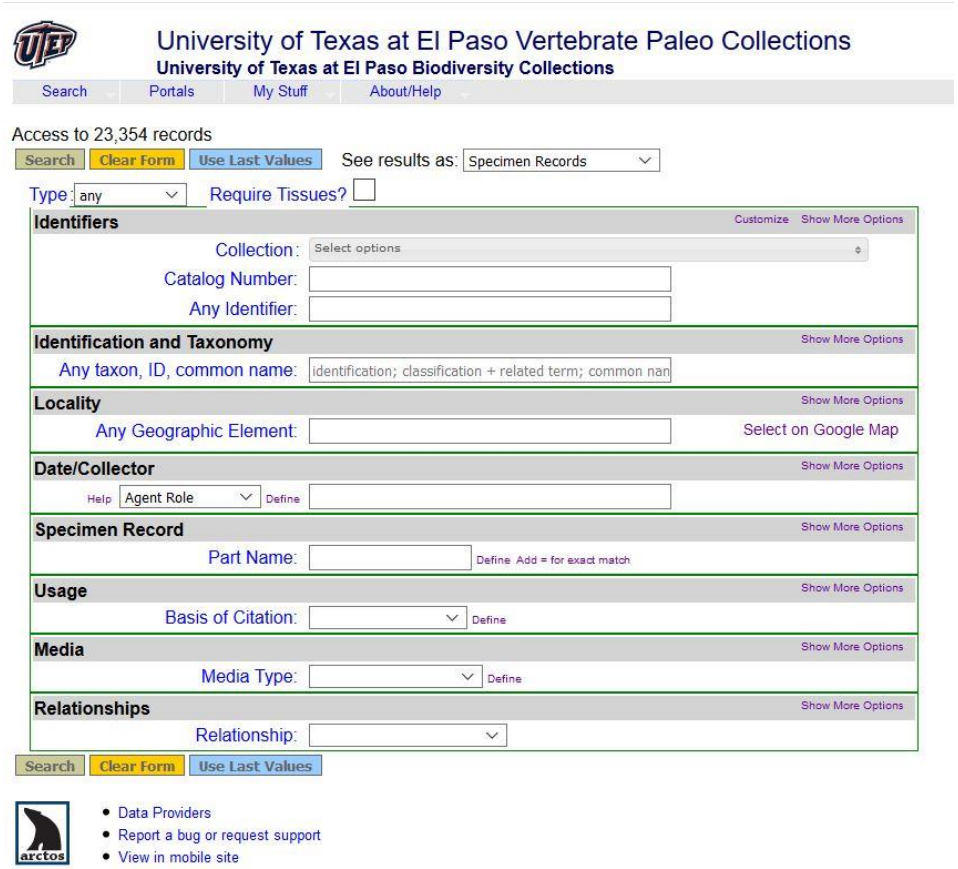
We've done a bit of thinking about this internally. Right now there are some data quality flags from iDigBio that are useful because they correct objectively incorrect data, like a mismatch between coordinates and country due to a missing sign. Others, like your Pacific islands example, are subjective to how data are stored in Arctos vs. other models. Many of the objective DQ tests would flag errors that we don't have because Arctos/Dusty also catches them (e.g. "April 31st" is a date that doesn't exist"). The subjective ones I don't think are worth our time we care about at this point, in particular because the DQ tests and methods iDigBio uses are in flux due to work being done in TDWG.

The TDWG Biodiversity Data Quality task group has a few iterations working on different aspects. One is trying to define a framework for what we even mean when we talk about data quality as a collections community. Another is getting all the aggregators, including iDigBio, to settle on a set of the *same* data quality tests to run on provider data and return flags for.

I don't actually think the flags are visibly negative enough to make users think "Arctos is broken." I would hope (although I guess hope is the operative word here), that people who are running analyses on or otherwise using aggregator data for something beyond browsing would notice that the flags are doing more standardizing than correcting, and that obviously different collections/databases use different but equally correct ways to say the same thing...

Arctos

One Collection's Data Flags



Arctos

Recordset

Search Recordset

23,354

UTEP Fossils (Arctos)

Specimen Records: 23,354 Media Records: 258

iDigBio Last Ingested Date: 2018-07-13

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	23354	100
geopoint_datum_missing	22963	98.326
dwc_datasetid_added	21141	90.524
dwc_parentnameusageid_added	21141	90.524
dwc_taxonid_added	21141	90.524
dwc_taxonomicstatus_added	21141	90.524
dwc_taxonrank_added	21141	90.524
gbif_canonicalname_added	21141	90.524
gbif_genericname_added	21141	90.524
gbif_taxon_corrected	21141	90.524
dwc_scientificnameauthorship_added	20615	88.272

Arctos

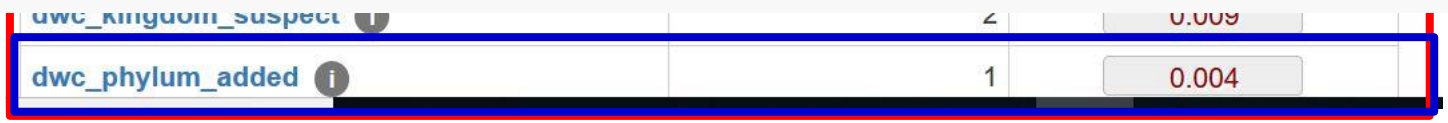
Data from source **Arctos** [Classifications] [Top]

last update: 2018-07-29 14:44:14.0
globalnames score not available
match type not available

author_text: A. Dietr.
display_name: *Picea* A. Dietr.
nomenclatural_code: ICBN
remark: Imported from ITIS 6 Feb 2007, added author text 7 Jul 2018 from ITIS
scientific_name: Picea
source_authority: ITIS
valid_catalog_term_fg: 1

Classification:

Plantae (kingdom) [more like this term] [including rank] [from this source]
Tracheophyta (phylum) [more like this term] [including rank] [from this source]
Pinopsida (class) [more like this term] [including rank] [from this source]
Pinales (order) [more like this term] [including rank] [from this source]
Pinaceae (family) [more like this term] [including rank] [from this source]
Picea (genus) [more like this term] [including rank] [from this source]



Arctos

Recordset

Search Recordset

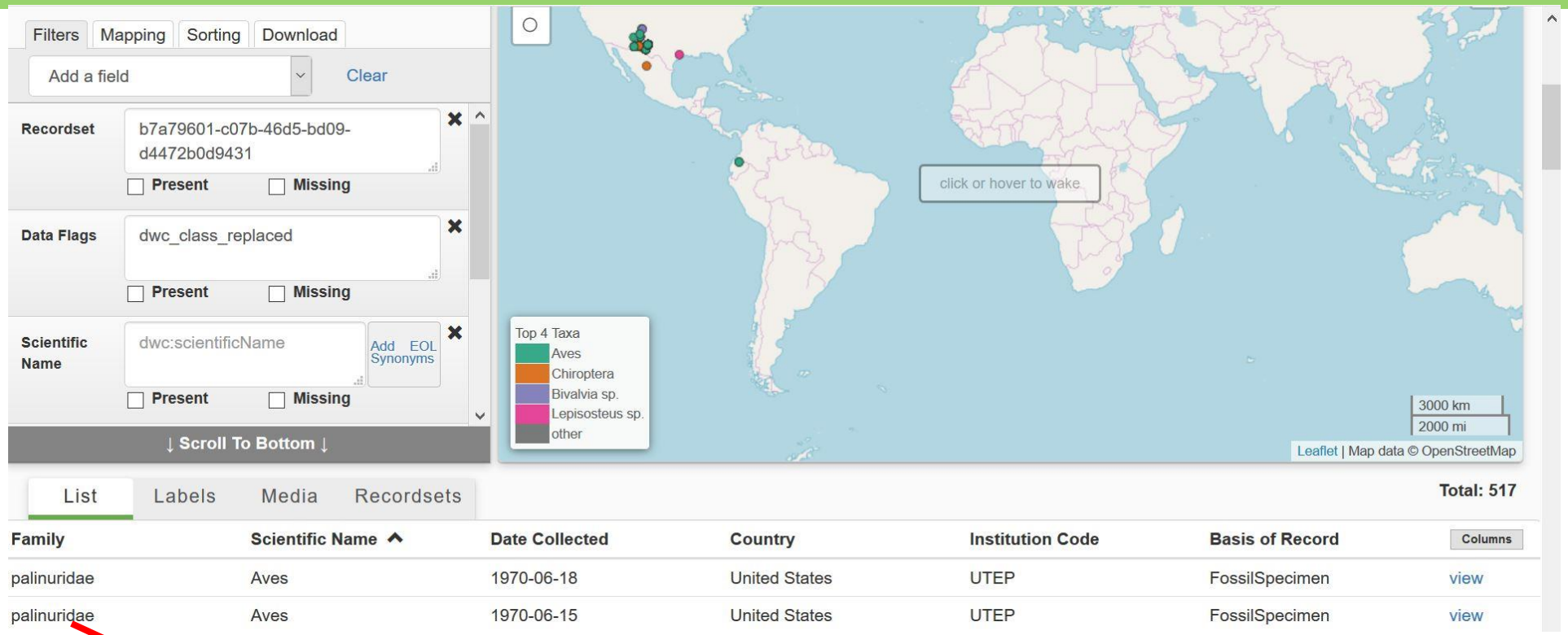
UTEP Fossils (Arctos)

Specimen Records: 23,354 Media Records: 258
iDigBio Last Ingested Date: 2018-07-13

taxon_match_failed	2274	9.737
dwc_specificepithet_added	2191	9.382
dwc_specificepithet_replaced	1666	7.134
dwc_family_replaced	1101	4.714
dwc_infraspecificepithet_added	1069	4.577
dwc_order_replaced	739	3.164
dwc_genus_replaced	645	2.762
dwc_family_added	524	2.244
dwc_genus_added	524	2.244
dwc_phylum_replaced	521	2.231
dwc_class_replaced	517	2.214
dwc_order_added	436	1.867
geopoint_low_precision	121	0.518
geopoint_datum_error	100	0.428
dwc_class_added	9	0.039
dwc_taxonremarks_added	4	0.017
dwc_kingdom_suspect	2	0.009
dwc_phylum_added	1	0.004

Arctos

Is Aves not a class?



Family=palinuridae

Arctos

What the what?

Arctos

The mystery deepens…

University of Texas at El Paso Vertebrate Paleo Collections
University of Texas at El Paso Biodiversity Collections

Search | Portals | My Stuff | About/Help

UTEP:ES:4-578
Aves
<< Return to results

Dry Cave, Bison Chamber
North America, United States, New Mexico, Eddy County
1971-- (1971-01-01 - 1971-12-31)

Identifications
Aves
Animalia; Chordata; Aves; Aves
Identified by Arthur H. Harris on 1971-12-31
Nature of ID: expert

Collector(s)
Arthur H.

Parts
Part Nam

Data from source **Arctos** [Classifications] [Top]

last update: 2017-06-01 00:35:18.0
globalnames score not available
match type not available

display_name: Aves
nomenclatural_code: ICZN
scientific_name: Aves
source_authority: ITIS
valid_catalog_term_fg: 1

Classification:

Animalia (kingdom) [more like this term] [including rank] [from this source]
Chordata (phylum) [more like this term] [including rank] [from this source]
Aves (class) [more like this term] [including rank] [from this source]

Genus = Avus

Data | Flags | Raw

Taxonomy

Scientific Name	Aves
Higher Classfication	Animalia; Chordata; ;
Kingdom	Animalia
Phylum	Chordata
Class	Aves
Order	decapoda
Family	palinuridae
Genus	avus
Nomenclatural Code	ICZN

Arctos

An answer, I think

Animalia > Chordata > Aves

Aves

Identification=scientific name=dwc:scientificName “Aves”

Taxon rank=dwc:taxonrank “class”

Arctos

What do biodiversity collection managers do all day?

Arctos

In My Backyard

We need more help at home!

- under-staffed and under-funded
- lack of experience/knowledge on the tech end
- assistance expressing the importance of the endeavor to administrators
- sustainability

Arctos

Thank You!

Photos courtesy of:

Arctos
Arthur H. Harris
Museum of Obsolete Media
Wikimedia Commons

All opinions expressed are my own. I would like to thank

Everyone in the **Arctos Working Group** for putting up with my multiple requests and issues on the Arctos GitHub. You all make my life so much richer!

Dr. Art Harris for being my iDigBio data quality flag guinea pig and partner in biological data crime.

Deb Paul for taking my questions seriously and making an honest attempt to answer them even while traveling.

Arctos