Establishing a New Framework for Paleontological Data Through an Evaluation of Current Data Sharing Practices

Holly Little ◆ littleh@si.edu
Department of Paleo Informatics Manager
Smithsonian National Museum of Natural History

Smithsonian Institution
NATIONAL MUSEUM of NATURAL HISTORY

TDWG 2018

Local: NMNH Paleo

662,353 (+-) Fossil Occurrence Records

Local: NMNH Paleo

662,353 (+-) Fossil Occurrence Records

1970-present: Creation of digital records

Local: NMNH Paleo

662,353 (+-) Fossil Occurrence Records

1970-present: Creation of digital records

Global

10,042,739 GBIF Fossil Occurrences

2012: Darwin Core (DwC) Standard

Local: NMNH Paleo

662,353 (+-) Fossil Occurrence Records

1970-present: Creation of digital records

Global

10,042,739 GBIF Fossil Occurrences

2012: Darwin Core (DwC) Standard

Local: NMNH Paleo

662,353 (+-) Fossil Occurrence Records

1970-present: Creation of digital records

Global

10,042,739 GBIF Fossil Occurrences

2012: Darwin Core (DwC) Standard

TDWG
PALEO IG

Local: NMNH Paleo

662,353 (+-) Fossil Occurrence Records

1970-present: Creation of digital records

Global

10,042,739 GBIF Fossil Occurrences

2012: Darwin Core (DwC) Standard

TDWG
PALEO IG

Local: NMNH Paleo

662,353 (+-) Fossil Occurrence Records

1970-present: Creation of digital records

- What data do we share?
- What isn't being shared? Why?
- Is the data clean and standardized?
- Are there other data formats or standards we could be sharing in?
  - ➢ New terms needed?

Global

10,042,739 GBIF Fossil Occurrences

2012: Darwin Core (DwC) Standard

TDWG
PALEO IG

NMNH Paleo Data

Issues and flags	Count
Taxon match none	153,080
Taxon match higherrank	97,043
Geodetic datum assumed WGS84	68,327
Taxon match fuzzy	15,263
Country derived from coordinates	12,692
Recorded date unlikely	3,187
Country invalid	1,317
Country coordinate mismatch	617
Recorded date invalid	207
Depth min/max swapped	92

81 DwC terms

type
references
institutionID
institutionCode
collectionCode
datasetName
basisOfRecord
occurrenceID
catalogNumber
recordNumber
recordedBy
individualCount
sex
lifeStage
preparations
associatedMedia
associatedSequences
otherCatalogNumbers
occurrenceRemarks
associatedOccurrences
fieldNumber
startDayOfYear
endDayOfYear
year
month
day
verbatimEventDate
habitat
fieldNotes
locationID
higherGeography
continent
waterBody
islandGroup
island
country
stateProvince
county
locality
minimumElevationInMeters
maximumElevationInMeters
verbatimElevation
minimumDepthInMeters
maximumDepthInMeters
decimalLatitude
decimalLongitude
geodeticDatum
coordinateUncertaintyInMeters
verbatimLatitude
verbatimLongitude
verbatimCoordinateSystem
georeferenceProtocol
georeferenceRemarks
earliestEraOrLowestErathem
latestEraOrHighestErathem
earliestPeriodOrLowestSystem
latestPeriodOrHighestSystem
earliestEpochOrLowestSeries
latestEpochOrHighestSeries
earliestAgeOrLowestStage
latestAgeOrHighestStage
group
Formation
verbatimDepth
member
identificatio
typeStatus
identifiedBy
scientificNar
higherClassi
kingdom
phylum
class
order
family
genus
subgenus
specificEpith
infraspecific
taxonRank
scientificNar

c8681cc2-9d0a-4c5f-b620-5c753abfe2bc

Micropaleo Data

How can we compare the micropaleo collections from the NMNH, NHM London, and AMNH?

- Giles Miller (NHM London)

Initial Query parameters:

- fossilSpecimen + Foraminifera
- fossilSpecimen + Ostracoda
- datasetName:PAL + Foraminifera
- Ostracoda
- Foraminifera
- datasetName: AMNH MICROFOSSIL

✔ institutionCode
✔ collectionCode
✔ basisOfRecord
✔ occurrenceID
✔ catalogNumber

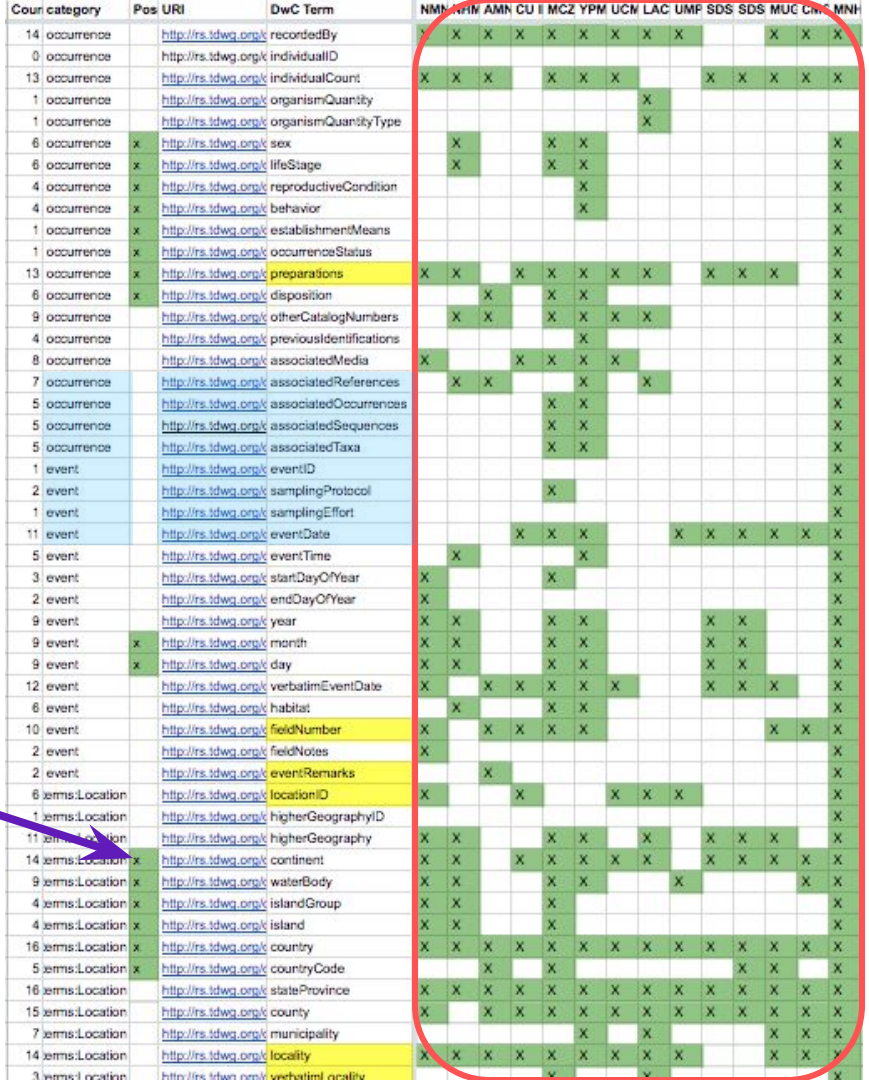
✘ ✔ Event terms
✘ ✔ Locality terms

101 DwC Terms used across 3 datasets

Testing

Key Elements:

- Testing for variation
- Generating list of terms and relationships between terms
- Highlighting points of interest
- Noting DwC terms that suggest controlled vocabulary



Coun	category	Pos	URI	DwC Term	NMN	MHM	AMN	CU	MCZ	YPM	UCM	LAC	UMF	SDS	SDS	MUC	CM	MNH
14	occurrence		http://rs.tdwg.org/c	recordedBy	X	X	X	X	X	X	X	X	X			X	X	X
0	occurrence		http://rs.tdwg.org/c	individualID														
13	occurrence		http://rs.tdwg.org/c	individualCount	X	X	X		X	X	X			X	X	X	X	X
1	occurrence		http://rs.tdwg.org/c	organismQuantity								X						
1	occurrence		http://rs.tdwg.org/c	organismQuantityType								X						
6	occurrence	x	http://rs.tdwg.org/c	sex		X			X	X								X
6	occurrence	x	http://rs.tdwg.org/c	lifeStage		X			X	X								X
4	occurrence	x	http://rs.tdwg.org/c	reproductiveCondition						X								X
4	occurrence	x	http://rs.tdwg.org/c	behavior						X								X
1	occurrence	x	http://rs.tdwg.org/c	establishmentMeans														X
1	occurrence	x	http://rs.tdwg.org/c	occurrenceStatus														X
13	occurrence	x	http://rs.tdwg.org/c	preparations	X	X		X	X	X	X	X		X	X	X		X
6	occurrence	x	http://rs.tdwg.org/c	disposition			X		X	X								X
9	occurrence		http://rs.tdwg.org/c	otherCatalogNumbers		X	X		X	X	X	X						X
4	occurrence		http://rs.tdwg.org/c	previousIdentifications						X								X
8	occurrence		http://rs.tdwg.org/c	associatedMedia	X			X	X	X	X							X
7	occurrence		http://rs.tdwg.org/c	associatedReferences		X	X			X		X						X
5	occurrence		http://rs.tdwg.org/c	associatedOccurrences					X	X								X
5	occurrence		http://rs.tdwg.org/c	associatedSequences					X	X								X
5	occurrence		http://rs.tdwg.org/c	associatedTaxa					X	X								X
1	event		http://rs.tdwg.org/c	eventID														X
2	event		http://rs.tdwg.org/c	samplingProtocol					X									X
1	event		http://rs.tdwg.org/c	samplingEffort														X
11	event		http://rs.tdwg.org/c	eventDate				X	X	X			X	X	X	X	X	X
5	event		http://rs.tdwg.org/c	eventTime		X				X								X
3	event		http://rs.tdwg.org/c	startDayOfYear	X				X									X
2	event		http://rs.tdwg.org/c	endDayOfYear	X													X
9	event		http://rs.tdwg.org/c	year	X	X			X	X				X	X			X
9	event	x	http://rs.tdwg.org/c	month	X	X			X	X				X	X			X
9	event	x	http://rs.tdwg.org/c	day	X	X			X	X				X	X			X
12	event		http://rs.tdwg.org/c	verbatimEventDate	X		X	X	X	X	X			X	X	X		X
6	event		http://rs.tdwg.org/c	habitat		X			X	X								X
10	event		http://rs.tdwg.org/c	fieldNumber	X		X	X	X	X						X	X	X
2	event		http://rs.tdwg.org/c	fieldNotes	X													X
2	event		http://rs.tdwg.org/c	eventRemarks			X											X
6	terms:Location		http://rs.tdwg.org/c	locationID	X			X			X	X	X					X
1	terms:Location		http://rs.tdwg.org/c	higherGeographyID														X
11	terms:Location		http://rs.tdwg.org/c	higherGeography	X	X			X	X		X		X	X	X		X
14	terms:Location	x	http://rs.tdwg.org/c	continent	X	X		X	X	X	X	X		X	X	X	X	X
9	terms:Location	x	http://rs.tdwg.org/c	waterBody	X	X			X	X			X				X	X
4	terms:Location	x	http://rs.tdwg.org/c	islandGroup	X	X			X									X
4	terms:Location	x	http://rs.tdwg.org/c	island	X	X			X									X
16	terms:Location	x	http://rs.tdwg.org/c	country	X	X	X	X	X	X	X	X	X	X	X	X	X	X
5	terms:Location	x	http://rs.tdwg.org/c	countryCode			X		X						X	X		X
16	terms:Location		http://rs.tdwg.org/c	stateProvince	X	X	X	X	X	X	X	X	X	X	X	X	X	X
15	terms:Location		http://rs.tdwg.org/c	county	X		X	X	X	X	X	X	X	X	X	X	X	X
7	terms:Location		http://rs.tdwg.org/c	municipality						X		X				X	X	X
14	terms:Location		http://rs.tdwg.org/c	locality	X	X	X	X	X	X	X	X	X			X	X	X
3	terms:Location		http://rs.tdwg.org/c	verbatimLocality														X

API Query

Datasets

Type: Occurrence

Metadata containing: Fossil

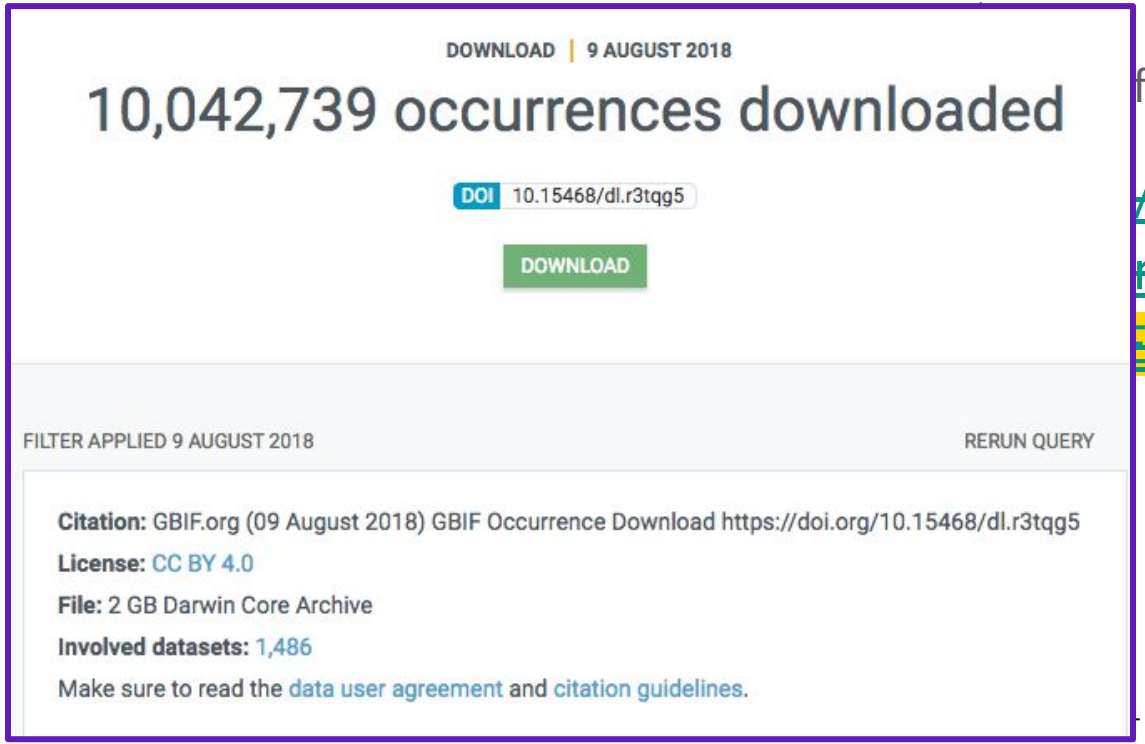
http://api.gbif.org/v1/dataset/search?q=fossil&type=occurrence

Occurrence Records

basisOfRecord: fossilSpecimen

https://api.gbif.org/v1/occurrence/search?basis_of_record=FOSSIL_SPECIMEN

~~API Query~~ Download



Occurrence Records

fRecord: fossilSpecimen

//api.gbif.org/v1/occurrenc
rch?basis_of_record=FOSSI
ECIMEN

Datasets

Total datasets:	1646
Tool:	Python*
Sources:	API Query (160)
	dataset list from occurrence download (1486)

*All scripts still need some debugging and cleaning

API Request

1. Query GBIF for datasets with ‘fossil’
2. Extract GBIF dataset Key
3. Query GBIF with dataset key
4. Find DwC-A endpoint
5. Request Zip file from endpoint
6. Unzip DwC-A
7. Extract metadata for dataset CSV report

Key elements:

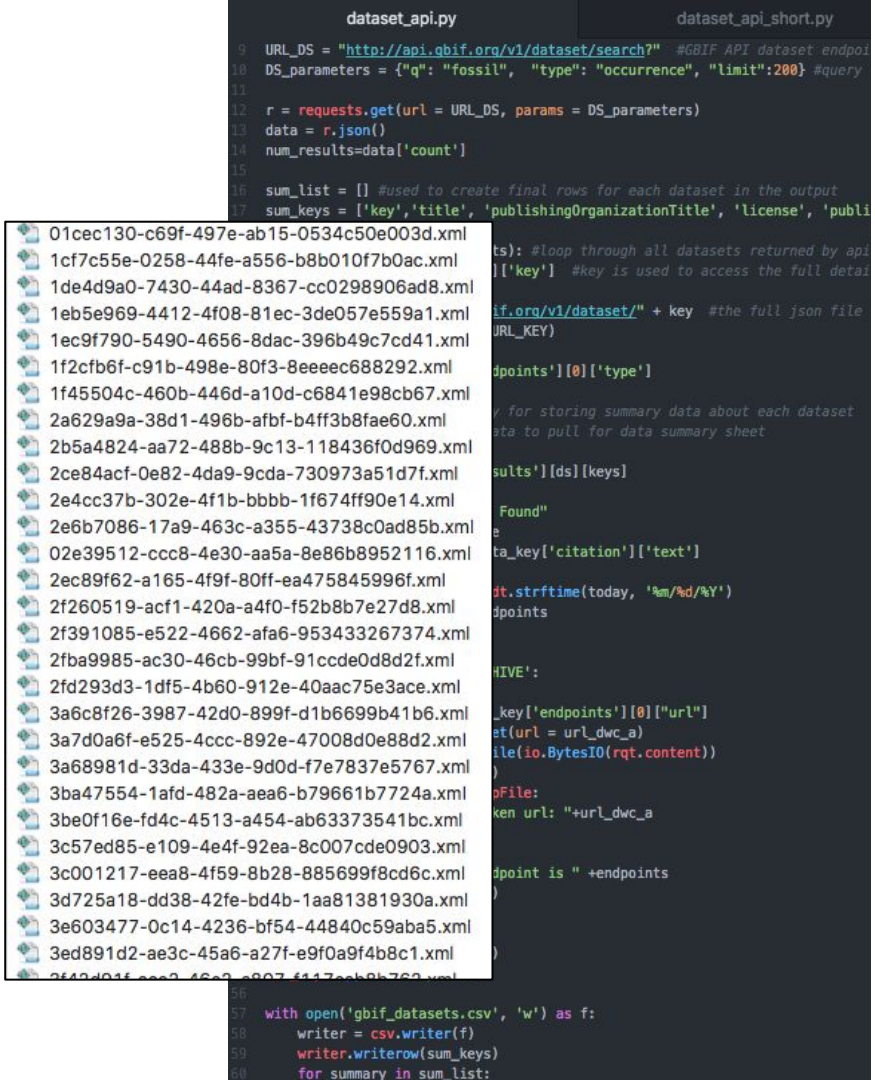
- Filtered down to endpoints that were tagged as DWC_ARCHIVE
- Account for inactive links or bad zip files

API Request

1. ~~Query GBIF for datasets with ‘fossil’~~
2. ~~Extract GBIF dataset Key~~
3. Query GBIF with dataset key
4. Find DwC-A endpoint
5. Request Zip file from endpoint
6. Unzip DwC-A
7. Extract metadata for dataset CSV report

Key elements:

- Filtered down to endpoints that were tagged as DWC_ARCHIVE
- Account for inactive links or bad zip files



Total datasets: 1476

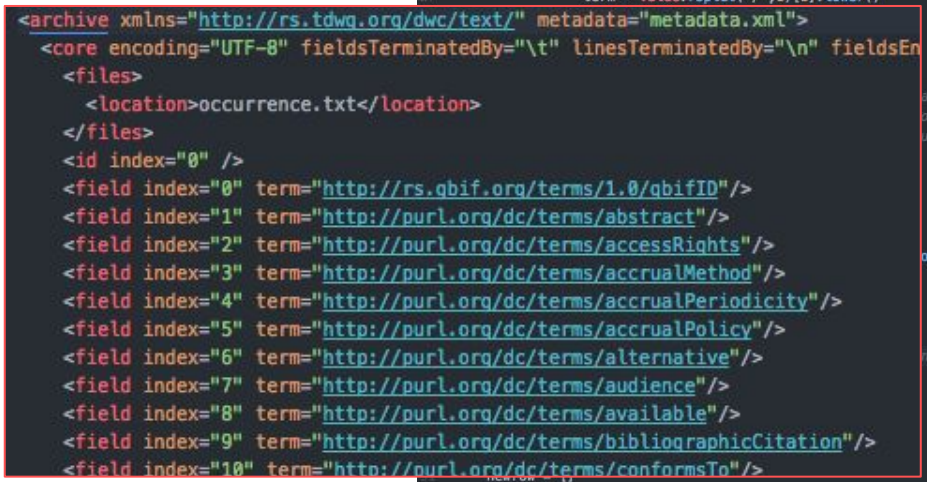
key	title	publis	license	publishingCc	recordCount	citation	accessDate	endpoint
7ca4f6dc-f762-11e1-a439-00145eb45e9a	(Table 2) Distribution of ice-rafted fossils i	PANG	http://creativecommons.org/licenses/by/4.0/	DE	48	Spiegler D (1988). (Table 2) Distribution of ice-raft	8/19/18	DWC_ARCHIVE
b275a4c1-9859-4f3c-8ead-d86dde820fbc	The fossil collection (F) of the MusÃ©um r	MNHN	http://creativecommons.org/licenses/by/4.0/	FR	Key Not Found	MNHN - Museum national d'Histoire naturelle (20	8/19/18	DWC_ARCHIVE
29334ea1-9b35-4370-9873-8d85ad56be6e	Fossil occurrence of planktonic foraminifer	PANG	http://creativecommons.org/licenses/by/4.0/	DE	18	Kinoshita M, Tobin H, Ashi J, Kimura G, Lallemant	8/19/18	DWC_ARCHIVE
815c3a6e-f762-11e1-a439-00145eb45e9a	(Table 1) Distribution of early late Eocene	PANG	http://creativecommons.org/licenses/by/4.0/	DE	2610	Nicora A, Premoli Silva I (1989). (Table 1) Distribut	8/19/18	DWC_ARCHIVE
680c12e3-6440-4559-8add-b263cb48f31a	(Table 2) Fossil record of foraminifera and	PANG	http://creativecommons.org/licenses/by/4.0/	DE	404	Woszidlo H (1961). (Table 2) Fossil record of forar	8/19/18	DWC_ARCHIVE
2e6b7086-17a9-463c-a355-43738c0ad85b	(Table 1) Fossil record of foraminifera and	PANG	http://creativecommons.org/licenses/by/4.0/	DE	353	Woszidlo H (1961). (Table 1) Fossil record of forar	8/19/18	DWC_ARCHIVE
d6618bc2-ce35-4c16-94e1-d58063736899	(Table 3) Minerals, and calcareous and silic	PANG	http://creativecommons.org/licenses/by/4.0/	DE	286	Gervais E (1986). (Table 3) Minerals, and calcareo	8/19/18	DWC_ARCHIVE
b201c8fb-0502-437f-ba77-8e092273328b	(Table 5) Record of mollusca and ostracod	PANG	http://creativecommons.org/licenses/by/4.0/	DE	120	Krienke H, Strahl J, Frenzel P, Keding E (1998). (Ta	8/19/18	DWC_ARCHIVE
81ca34f6-f762-11e1-a439-00145eb45e9a	(Table 2) Distribution of early late Eocene	PANG	http://creativecommons.org/licenses/by/4.0/	DE	36	Nicora A, Premoli Silva I (1989). (Table 2) Distribut	8/19/18	DWC_ARCHIVE
f6a07b42-1d2c-11e2-8fd4-00145eb45e9a	Collection Paleontology - GPIT	Sencke	http://creativecommons.org/licenses/by/4.0/	DE	28350	Senckenberg. Collection Paleontology - GPIT. Occ	8/19/18	BIOCASE
1aecc602-8847-4980-b493-bb21bb4aace5	PalÃ¤ontologie Marburg	Sencke	http://creativecommons.org/licenses/by/4.0/	DE	5166	Senckenberg. PalÃ¤ontologie Marburg. Occurren	8/19/18	BIOCASE
40275fc6-3e71-4d76-aafe-276bdeefcd86	SDSM Botany and Paleobotany Collections	South	http://creativecommons.org/publicdomain/z	US	Key Not Found	Hess G, Shelton S (2018). SDSM Botany and Paleo	8/19/18	DWC_ARCHIVE
71fa736a-f762-11e1-a439-00145eb45e9a	MfN - Glacial erratics	Museu	http://creativecommons.org/licenses/by/4.0/	DE	2	Museum fÃ¼r Naturkunde Berlin. MfN - Glacial er	8/19/18	BIOCASE
21a0cb45-f0ec-4901-a868-cd45abe74601	Canadian Museum of Nature Palynology C	Canad	http://creativecommons.org/licenses/by/4.0/	CA	14568	Shepherd K, Shorthouse D (2018). Canadian Muse	8/19/18	DWC_ARCHIVE
53e9eca5-6831-4a43-a109-130feecfcd9c	Cleveland Museum of Natural History Inve	Clevel	http://creativecommons.org/publicdomain/z	US	20585	Dunn D (2018). Cleveland Museum of Natural Hist	8/19/18	DWC_ARCHIVE
c8681cc2-9d0a-4c5f-b620-5c753abfe2bc	NMNH Paleobiology Specimen Records	Natior	http://creativecommons.org/publicdomain/z	US	662355	Orrell T, Hollowell T (2018). NMNH Paleobiology S	8/19/18	DWC_ARCHIVE
6720aee6-2aad-446d-bb97-ba009d1b5666	CM Vertebrate Paleontology Collection	Carne	http://creativecommons.org/publicdomain/z	US	79254	Henrici A (2016). CM Vertebrate Paleontology Col	8/19/18	DWC_ARCHIVE
83216388-f762-11e1-a439-00145eb45e9a	ColecciÃ³n PaleontolÃ³gica Invertebrados	CCT C	http://creativecommons.org/licenses/by/4.0/	AR	190	CCT CONICET-CENPAT Centro CientÃ-fico TecnolÃ	8/19/18	DIGIR
361d98ca-ba9d-4b85-a64f-7db8fb35c6a9	CMC Cincinnati Museum Center Invertebr	Cincin	http://creativecommons.org/licenses/by-nc/4	US	60165	Hunda B, Kling A, Storrs G (2016). CMC Cincinnati	8/19/18	DWC_ARCHIVE
95da13d0-f762-11e1-a439-00145eb45e9a	ELM paleontology	Geoco	http://creativecommons.org/licenses/by-nc/4	EE	24765	Geocollections of Estonia. ELM paleontology. Occ	8/19/18	BIOCASE
8130e5c6-f762-11e1-a439-00145eb45e9a	IG TUT paleontology	Geoco	http://creativecommons.org/licenses/by-nc/4	EE	110964	Geocollections of Estonia. IG TUT paleontology. O	8/19/18	BIOCASE
8f7e3c45-4d76-4982-9478-651439e0cd4b	Natural History Museum, University of Tar	PlutoF	http://creativecommons.org/licenses/by-nc/4	EE	Key Not Found	PlutoF (2018). Natural History Museum, Universit	8/19/18	DWC_ARCHIVE
561367c9-576b-4aa8-8fc9-1981155b11af	Museo Argentino de Ciencias Naturales "B	Museo	http://creativecommons.org/licenses/by/4.0/	AR	Key Not Found	Gutierrez P R, RodrÃ-guez D (2018). Museo Argen	8/19/18	DWC_ARCHIVE
85530bf4-67af-4c8e-a176-0ea3c322d9fc	Museo Argentino de Ciencias Naturales "B	Museo	http://creativecommons.org/licenses/by/4.0/	AR	Key Not Found	Genise J, RodrÃ-guez D (2018). Museo Argentino	8/19/18	DWC_ARCHIVE
f5783452-6ff5-4bef-8f30-9adfeba81bd7	Paleontological Research Institution Collec	Paleor	http://creativecommons.org/publicdomain/z	US	19541	Skibinski L (2018). Paleontological Research Instit	8/19/18	DWC_ARCHIVE
e1e16cf0-ada2-11e2-8fbc-00145eb45e9a	Neptune Deep-Sea Microfossil Occurrence	Museu	http://creativecommons.org/licenses/by/4.0/	DE	500808	Museum fÃ¼r Naturkunde Berlin. Neptune Deep-	8/19/18	BIOCASE
3c001217-eea8-4f59-8b28-885699f8cd6c	Ohio University Invertebrate Paleontology	Ohio U	http://creativecommons.org/publicdomain/z	US	2412	Stigall A (2016). Ohio University Invertebrate Pale	8/19/18	DWC_ARCHIVE
bea28c6b-4282-4e0e-894d-7c65d050ffa9	NCSM Invertebrate Paleontology Collectio	North	http://creativecommons.org/publicdomain/z	US	11877	Norton B (2016). NCSM Invertebrate Paleontology	8/19/18	DWC_ARCHIVE
0ec927cf-325a-4d63-9499-d721c734463a	LACM Entomology Collection	Natur	http://creativecommons.org/publicdomain/z	US	184364	Mertz B (2018). LACM Entomology Collection. Ver	8/19/18	DWC_ARCHIVE
b25c3e4d-4742-4d1a-a8cb-4b7648bace74	CSIRO Survey TT200801 - ROV Jason cruise	CSIRO	http://creativecommons.org/licenses/by/4.0/	AU	1353	Althaus F, Watts D (2017). CSIRO Survey TT20080	8/19/18	DWC_ARCHIVE
b78dc638-1dde-43b4-a4cb-1e7e8ba2f185	ColecciÃ³n de moldes endocraneanos del	CCT C	http://creativecommons.org/licenses/by/4.0/	AR	146	Dozo T (2015). ColecciÃ³n de moldes endocranea	8/19/18	DWC_ARCHIVE
ee789ae4-ef51-4ff2-931b-bc61b2dbe40e	Museo Argentino de Ciencias Naturales "B	Museo	http://creativecommons.org/licenses/by/4.0/	AR	Key Not Found	Kramarz A, RodrÃ-guez D (2018). Museo Argentin	8/19/18	DWC_ARCHIVE
eb6a3fcc-07e8-4cea-b21c-93382834f490	Collection scientifique et pÃ©dagogique d	Ecole	http://creativecommons.org/licenses/by/4.0/	FR	581	Moussus J (2014). Collection scientifique et pÃ©d	8/19/18	DWC_ARCHIVE
048d1f67-f86c-441d-b6f9-04b1b464e146	Museo Argentino de Ciencias Naturales "B	Museo	http://creativecommons.org/licenses/by/4.0/	AR	Key Not Found	Kramarz A, RodrÃ-guez D (2018). Museo Argentin	8/19/18	DWC_ARCHIVE
463a3bd9-a9bf-4667-ae34-44b023307367	Museo Argentino de Ciencias Naturales "B	Museo	http://creativecommons.org/licenses/by/4.0/	AR	Key Not Found	Totah V, RodrÃ-guez D (2018). Museo Argentino	8/19/18	DWC_ARCHIVE
7814596d-f762-432d-abda-a2e5ad64df92	Museo Argentino de Ciencias Naturales "B	Museo	http://creativecommons.org/licenses/by/4.0/	AR	Key Not Found	Del Fuevo G, RodrÃ-guez D (2018). Museo Argent	8/19/18	DWC_ARCHIVE

Meta.xml

1. Open meta.xml file for each DwC-A
2. Read terms and assign “X” for term presence
3. Count total number of terms
4. Count number of extensions
5. Record extension names
6. Append results to CSV

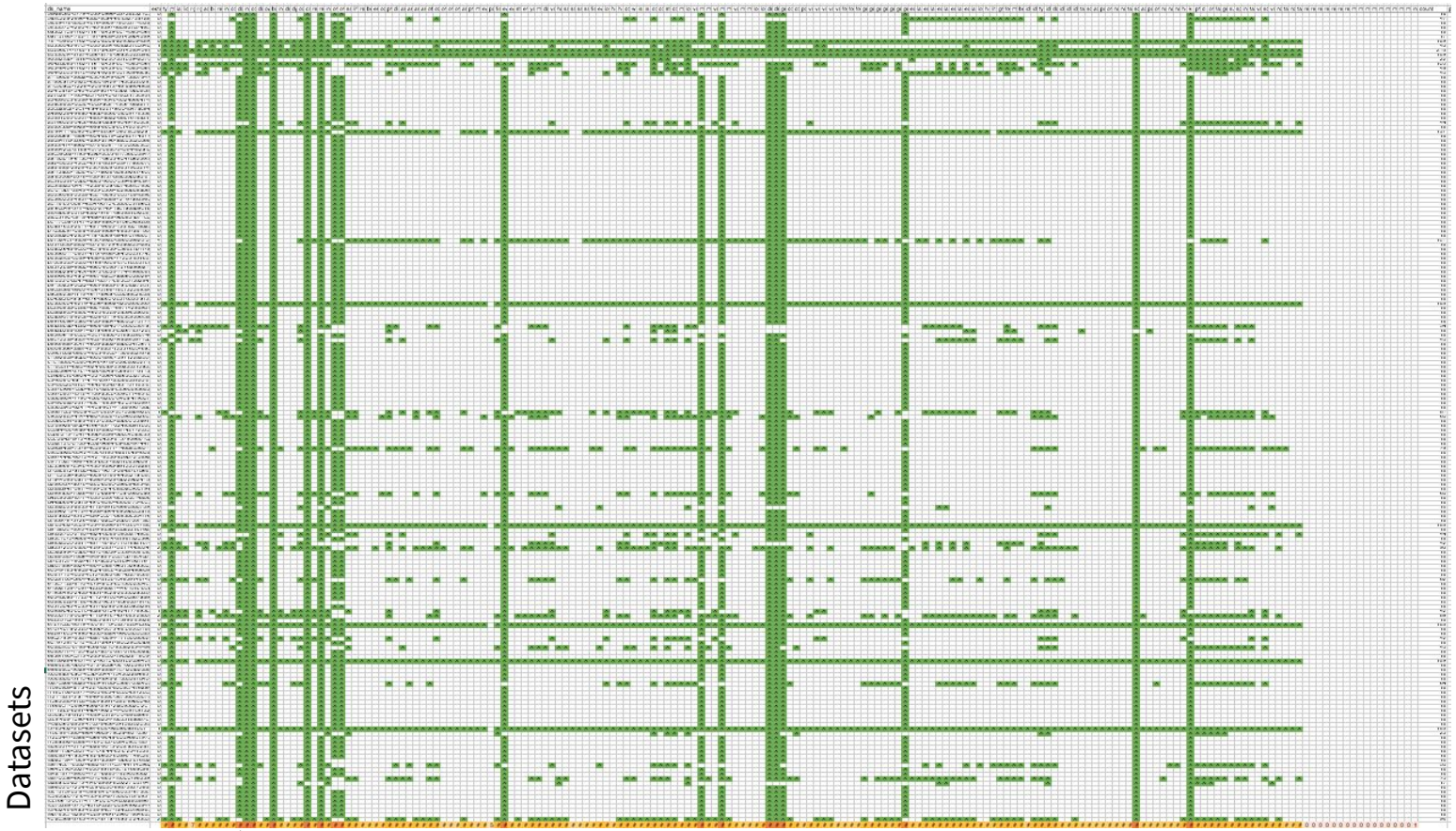
Key elements:

- Some terms not checked
- Lowercase all terms
- Does not account for poor formatting
- Removed ~900 rows for variations on the same dataset



Total datasets: 244

Average # terms: 32



Datasets

DwC Terms (In all charts terms are ordered by class and as listed in DwC specifications)

occurrence.txt (~~.tab~~)

1. Load .txt file into dataframe
2. Count number of usages per term
3. Calculate percentage
4. Append results to CSV

Key elements:

- Cannot process .tab files
- Lowercase all terms
- Skip lines with poorly formatted data (negligible)
- Probably includes none paleo data
- Memory errors
    - Run datasets in smaller batches

dwc-a_r.py | dwc-a_xml.py

```
import pandas as pd
import numpy as np
import os
import csv


directory_list = []
path = r"/Volumes/HL_64_USB3/Paleo DwC/"
for child in os.listdir(path):
    if os.path.isdir(child):
        dirpath = os.path.join(path, child)
        directory_list.append(dirpath)

usages = []
for folder in directory_list:
    print(folder)
    usage={}
    try:
        occurrence = pd.read_table(folder+"/occurrence.txt", dtype = 'object', error
        row_total = occurrence.shape[0]
        print(row_total)

        usage['total_count'] = row_total
        usage['set_name'] = os.path.basename(folder)     #metadata to identify each
        usage['institution_Code'] = occurrence['institutionCode'][1]

        for column in occurrence:     #loop to count number of times a term is used
            counts = occurrence[column].count()
            percentage = counts/row_total*100
#            usage[column] = counts
            usage[column.lower()] = percentage

        usages.append(usage)

    except FileNotFoundError:
        usage['set_name'] = os.path.basename(folder)
        usage['error'] = 'File Not Found'
        usages.append(usage)

keys = []
keys = ['set_name','institution_Code', 'error', 'total_count']
with open('dwc_terms.csv', 'r') as f:
    for line in f:
        keys.append(line.strip().lower())

#with open('occurrence_count.csv', 'a', newline ='') as f:
with open('occurrence_percent.csv', 'a', newline ='') as f:
    writer = csv.writer(f)
#    writer.writerow(keys)
    for usage in usages:
        row = [usage.get(key,"") for key in keys]
        writer.writerow(row)
```

Total datasets: 101

Percentages

Total datasets: 101

Percentages

Event Location GeoRef Geo Context Taxonomy

Presence

Top 10% (~200 usages)
modified
datasetID
institutionCode
collectionCode
basisOfRecord
catalogNumber
recordedBy
organismQuantity
organismQuantityType
eventDate
minimumElevationInMeters
minimumDepthInMeters
decimalLatitude
decimalLongitude
geodeticDatum
geologicalContextID
scientificName
kingdom

Percentages

Top 10% (>75%)
type
modified
language
institutionCode
collectionCode
basisOfRecord
occurrenceID
catalogNumber
countryCode
stateProvince
verbatimLocality
minimumElevationInMeters
MinimumDepthInMeters
namePublishedIn
subgenus
relatedResourceID

Counts

Top 10% (>950,000)
modified
institutionCode
collectionCode
basisOfRecord
occurrenceID
catalogNumber
stateProvince
verbatimLocality
minimumElevationInMeters
MinimumDepthInMeters
namePublishedIn
subgenus
specificEpithet
infraspecificEpithet
taxonRank
verbatimTaxonRank
scientificNameAuthorship

Presence

Top 10% (~200 usages)

modified
datasetID
institutionCode
collectionCode
basisOfRecord
catalogNumber
recordedBy
organismQuantity
organismQuantityType
eventDate
minimumElevationInMeters
minimumDepthInMeters
decimalLatitude
decimalLongitude
geodeticDatum
geologicalContextID
scientificName
kingdom

Percentages

Top 10% (>75%)

type
modified
language
institutionCode
collectionCode
basisOfRecord
occurrenceID
catalogNumber
countryCode
stateProvince
verbatimLocality
minimumElevationInMeters
MinimumDepthInMeters
namePublishedIn
subgenus
relatedResourceID

Percent with Fossil filter

Top 10% (of 6,152,144 records)

type
modified
institutionCode
collectionCode
basisOfRecord
occurrenceID
catalogNumber
stateProvince
verbatimLocality
minimumElevationInMeters
MinimumDepthInMeters
namePublishedIn
Subgenus
specificEpithet
infraspecificEpithet
scientificNameAuthorship

Occurrence Records

Occurrences:	10,042,739
Tool:	SQL* BigQuery*
Sources:	Occurrence download

**Processing by Luis Villanueva, SI DPO*

Usage Counts

```
filename="columns_verbatim.csv"
while read -r line
do
    echo "$line"
    echo "$line" >> results_verbatim.txt

    echo "insert into gbif_paleo_20180718.gbif_results_verbatim (colname, querytype, value) (select '${line//[$'\t\r\n ']}', 'is_null', cast(count(gbifID) as string) FROM gbif_paleo_20180718.gbif_verbatim WH
    bq query --use_legacy_sql=false  "insert into gbif_paleo_20180718.gbif_results_verbatim (colname, querytype, value) (select '${line//[$'\t\r\n ']}', 'is_null', cast(count(gbifID) as string) FROM gbif_pal
    echo "insert into gbif_paleo_20180718.gbif_results_verbatim (colname, querytype, value) (select '${line//[$'\t\r\n ']}', 'no_distinct', cast(count(distinct ${line//[$'\t\r\n ']}) as string) FROM gbif_pal
    bq query --use_legacy_sql=false  "insert into gbif_paleo_20180718.gbif_results_verbatim (colname, querytype, value) (select '${line//[$'\t\r\n ']}', 'no_distinct', cast(count(distinct ${line//[$'\t\r\n '

done < "$filename"
```

```
eventID
insert into gbif_paleo_20180718.gbif_results_verbatim (colname, querytype, value) (select 'eventID', 'is_null', cast(count(gbifID) as string) FROM gbif_paleo_20180718.g
insert into gbif_paleo_20180718.gbif_results_verbatim (colname, querytype, value) (select 'eventID', 'no_distinct', cast(count(distinct eventID) as string) FROM gbif_pa
parentEventID
insert into gbif_paleo_20180718.gbif_results_verbatim (colname, querytype, value) (select 'parentEventID', 'is_null', cast(count(gbifID) as string) FROM gbif_paleo_2018
insert into gbif_paleo_20180718.gbif_results_verbatim (colname, querytype, value) (select 'parentEventID', 'no_distinct', cast(count(distinct parentEventID) as string)
fieldNumber
insert into gbif_paleo_20180718.gbif_results_verbatim (colname, querytype, value) (select 'fieldNumber', 'is_null', cast(count(gbifID) as string) FROM gbif_paleo_201807
insert into gbif_paleo_20180718.gbif_results_verbatim (colname, querytype, value) (select 'fieldNumber', 'no_distinct', cast(count(distinct fieldNumber) as string) FRO
eventDate
insert into gbif_paleo_20180718.gbif_results_verbatim (colname, querytype, value) (select 'eventDate', 'is_null', cast(count(gbifID) as string) FROM gbif_paleo_20180718
insert into gbif_paleo_20180718.gbif_results_verbatim (colname, querytype, value) (select 'eventDate', 'no_distinct', cast(count(distinct eventDate) as string) FROM gbi
```

Top 10%

basisOfRecord	is_null	0	0.00%	no_dist	7
collectionCode	is_null	1,852	0.02%	no_dist	182,414
institutionCode	is_null	432,376	3.95%	no_dist	119
scientificName	is_null	609,520	5.57%	no_dist	797,610
identifier	is_null	1,078,133	9.85%	no_dist	7,542,714
catalogNumber	is_null	2,275,400	20.78%	no_dist	6,321,787
country	is_null	2,797,005	25.55%	no_dist	1,644
kingdom	is_null	2,891,281	26.41%	no_dist	52
occurrenceID	is_null	3,298,333	30.13%	no_dist	6,734,061
recordedBy	is_null	3,783,676	34.56%	no_dist	80,189
modified	is_null	3,810,201	34.80%	no_dist	662,352
genus	is_null	4,196,477	38.33%	no_dist	108,006
decimalLatitude	is_null	4,488,154	41.00%	no_dist	117,132
decimalLongitude	is_null	4,488,191	41.00%	no_dist	127,247
class	is_null	4,530,641	41.38%	no_dist	845
phylum	is_null	4,579,874	41.83%	no_dist	392
continent	is_null	4,777,000	43.63%	no_dist	95
stateProvince	is_null	4,780,674	43.67%	no_dist	8,055
locality	is_null	4,968,186	45.38%	no_dist	353,130
geodeticDatum	is_null	5,153,474	47.07%	no_dist	66
order	is_null	5,188,502	47.39%	no_dist	3,173
family	is_null	5,225,724	47.73%	no_dist	15,098
type	is_null	5,553,827	50.73%	no_dist	8
specificEpithet	is_null	5,795,788	52.94%	no_dist	149,375
higherGeography	is_null	6,120,647	55.91%	no_dist	32,384
language	is_null	6,254,868	57.13%	no_dist	7
formation	is_null	6,520,756	59.56%	no_dist	31,500
higherClassification	is_null	6,673,749	60.96%	no_dist	123,029
nomenclaturalCode	is_null	6,833,749	62.42%	no_dist	6
taxonRank	is_null	6,868,427	62.74%	no_dist	93
county	is_null	6,975,852	63.72%	no_dist	11,063
datasetName	is_null	7,082,686	64.69%	no_dist	176,268
license	is_null	7,092,843	64.79%	no_dist	106
accessRights	is_null	7,150,863	65.32%	no_dist	9
earliestPeriodOrLowe	is_null	7,354,179	67.17%	no_dist	473
earliestAgeOrLowestS	is_null	7,671,117	70.07%	no_dist	2,576
individualCount	is_null	7,921,337	72.35%	no_dist	865
preparations	is_null	7,933,821	72.47%	no_dist	87,058
verbatimLatitude	is_null	7,944,984	72.57%	no_dist	62,871
verbatimLongitude	is_null	7,944,984	72.57%	no_dist	74,637
previousIdentification	is_null	8,004,903	73.12%	no_dist	438,445
eventDate	is_null	8,015,021	73.21%	no_dist	27,157
earliestEpochOrLowe	is_null	8,081,951	73.82%	no_dist	962
dynamicProperties	is_null	8,178,769	74.71%	no_dist	190,553
verbatimLocality	is_null	8,234,354	75.21%	no_dist	49,462
countryCode	is_null	8,262,288	75.47%	no_dist	450
georeferenceRemarks	is_null	8,286,071	75.69%	no_dist	14,702
verbatimTaxonRank	is_null	8,302,772	75.84%	no_dist	29
associatedReferences	is_null	8,334,040	76.12%	no_dist	56,143
lithostratigraphicTerm	is_null	8,366,550	76.42%	no_dist	82,766
occurrenceStatus	is_null	8,393,663	76.67%	no_dist	5
collectionID	is_null	8,449,259	77.18%	no_dist	180,652
georeferenceVerificat	is_null	8,594,634	78.50%	no_dist	11
taxonID	is_null	8,649,671	79.01%	no_dist	185,664
samplingProtocol	is_null	8,701,708	79.48%	no_dist	3,036
coordinatePrecision	is_null	8,714,138	79.60%	no_dist	30
namePublishedInYear	is_null	8,721,800	79.67%	no_dist	217
taxonConceptID	is_null	8,794,612	80.33%	no_dist	157,772
institutionID	is_null	8,954,946	81.80%	no_dist	23
occurrenceRemarks	is_null	8,990,319	82.12%	no_dist	772,952
earliestEraOrLowestE	is_null	9,035,302	82.53%	no_dist	100
rightsHolder	is_null	9,252,032	84.51%	no_dist	25
coordinateUncertaint	is_null	9,252,925	84.52%	no_dist	10,163
datasetID	is_null	9,331,943	85.24%	no_dist	1,329
references	is_null	9,397,213	85.84%	no_dist	1,543,006
locationID	is_null	9,419,653	86.04%	no_dist	119,272
group	is_null	9,421,710	86.06%	no_dist	3,534
year	is_null	9,510,404	86.87%	no_dist	315
identifiedBy	is_null	9,580,177	87.51%	no_dist	10,404
latestPeriodOrHighest	is_null	9,746,247	89.02%	no_dist	109

Distinct Values

```
filename="distinct_terms.txt"
while read -r line
do
    echo "$line"
    echo "$line" >> results_verbatim.txt

    echo "insert into gbif_paleo_20180718.gbif_results_verbatim_counts_datasets (colname, value, no_records, datasetName, title, InstitutionC
    bq query --use_legacy_sql=false  "insert into gbif_paleo_20180718.gbif_results_verbatim_counts_datasets (colname, value, no_records, datas
done < "$filename"
```

type
language
license
basisOfRecord
sex
lifeStage
reproductiveCondition
behavior
establishmentMeans
occurrenceStatus
preparations
disposition
continent
waterBody
islandGroup
island
countryCode
verbatimSRS
geodeticDatum
georeferenceVerificationStatus
typeStatus
taxonRank
nomenclaturalCode
taxonomicStatus

Controlled Vocabularies

colname	querytype	value	% of rows		
basisOfRecord	is_null	0	0.00%	no_distinct	7
country	is_null	2,797,005	25.55%	no_distinct	1,644
kingdom	is_null	2,891,281	26.41%	no_distinct	52
class	is_null	4,530,641	41.38%	no_distinct	845
phylum	is_null	4,579,874	41.83%	no_distinct	392
continent	is_null	4,777,000	43.63%	no_distinct	95
stateProvince	is_null	4,780,674	43.67%	no_distinct	8,055
geodeticDatum	is_null	5,153,474	47.07%	no_distinct	66
family	is_null	5,225,724	47.73%	no_distinct	15,098
type	is_null	5,553,827	50.73%	no_distinct	8
taxonRank	is_null	6,868,427	62.74%	no_distinct	93
county	is_null	6,975,852	63.72%	no_distinct	11,063
preparations	is_null	7,933,821	72.47%	no_distinct	87,058
earliestEpochOrLowestSeries	is_null	8,081,951	73.82%	no_distinct	962

Preparations: 87,058 values

Important Paleo information, but we use it for multiple purposes:

- Material Type
- Prep work done
  - By % or action
- Anatomy/Morphology
- Internal notations of work done

value	no_records	datasetName
fossilized	421324	
default - 1	358025	
unknown (fossil)	183683	
PREP - 1	139514	
fossil	86342	cuinvert
Fossil - 1	84565	University of Kansas Biodive
Dry - 1	63153	
Permineralization - 1	61484	University of Kansas Biodive
Cataloged - 1	55998	
fossil	49369	Fossil Vertebrate
fossil - 1	44262	
PREP - 2	40680	
Secondary microslides	33016	NMNH Paleobiology
shell	31780	
fossil - 1	28795	
Microslide	23021	NMNH Paleobiology
PREP - 3	18047	
Primary microslides	17088	NMNH Paleobiology
Compression/Impression - 1	16514	University of Kansas Biodive
100% prepped - 2	1741	
shell	1727	
tooth, molar	1721	
Dry - 5	1720	
humerus	1686	
Embedded	698	
SHELL FRAGS	697	
CALCANEUM	697	
?	697	
Tooth; incomplete	693	
tooth fragment	685	
HÃ²mer|D|	685	
Dry - 20	679	

Presence

Top 10% (~200 usages)

modified
datasetID
institutionCode
collectionCode
basisOfRecord
catalogNumber
recordedBy
organismQuantity
organismQuantityType
eventDate
minimumElevationInMeters
minimumDepthInMeters
decimalLatitude
decimalLongitude
geodeticDatum
geologicalContextID
scientificName
kingdom

Percentages

Top 10% (>75%)

type
modified
language
institutionCode
collectionCode
basisOfRecord
occurrenceID
catalogNumber
countryCode
stateProvince
verbatimLocality
minimumElevationInMeters
MinimumDepthInMeters
namePublishedIn
subgenus
relatedResourceID

Counts

Top 10% (>950,000)

modified
institutionCode
collectionCode
basisOfRecord
occurrenceID
catalogNumber
stateProvince
verbatimLocality
minimumElevationInMeters
MinimumDepthInMeters
namePublishedIn
subgenus
specificEpithet
infraspecificEpithet
taxonRank
verbatimTaxonRank
scientificNameAuthorship

Occurrence Counts

Top 10%

basisOfRecord
collectionCode
institutionCode
scientificName
identifier
catalogNumber
country
kingdom
occurrenceID
recordedBy
modified
genus
decimalLatitude
decimalLongitude
class
phylum
continent
stateProvince
Locality
geodeticDatum
order

Problems

higherTaxonomy - only 30 datasets or 40% of records - Important for Paleo data (CLADES)

taxonRank - only 41 datasets or 38% of records (verbatimTaxonRank = 25%) with 93 distinct values

Preparations - only 47 datasets or 30% of records with 87,058 distinct values
We talk about it a lot, but aren't serving a lot of data and the data we do serve is very variable. Is it because we're not sure what to share?

Extensions - Barely using extensions (14 with multimedia, only 9 other extensions across all datasets)

Fixing NMNH Paleo

DwC Terms	USN	USNM p	All datasets percentage			All occurences		
eventDate			99.499221	is_	8,015,021	73.21%	no_	27,157
eventTime			#DIV/0!	is_	10,903,619	99.60%	no_	690
startDayOfYear	X	0.323	47.408504	is_	10,590,617	96.74%	no_	367
endDayOfYear	X		64.269201	is_	10,746,718	98.16%	no_	366
year	X		53.19237	is_	9,510,404	86.87%	no_	315
month	X	21.57	28.515609	is_	10,097,654	92.23%	no_	30
day	X	21.57	34.120706	is_	4,488,154	41.00%	no_	117,132
verbatimEventDate	X	37.61	59.913649	is_	9,995,065	91.30%	no_	33,090
habitat	X	21.69	48.135115	is_	10,945,077	99.97%	no_	73
fieldNumber	X	18.49	45.138677	is_	10,454,413	95.49%	no_	147,734
fieldNotes	X	36.94	47.701501	is_	10,947,675	100.00%	no_	1
eventRemarks			43.295624	is_	10,915,548	99.70%	no_	2,090
locationID	X		58.23435	is_	9,419,653	86.04%	no_	119,272
higherGeographyID			#DIV/0!	is_	10,945,961	99.98%	no_	3
higherGeography	X		#DIV/0!	is_	6,120,647	55.91%	no_	32,384
continent	X		0.9233132	is_	4,777,000	43.63%	no_	95
waterBody	X		66.666743	is_	10,844,516	99.06%	no_	282
islandGroup	X		29.616996	is_	10,902,702	99.59%	no_	65
island	X	50.77	77.392119	is_	10,897,567	99.54%	no_	319
country	X		100	is_	2,797,005	25.55%	no_	1,644
countryCode		78.29	94.347025	is_	8,262,288	75.47%	no_	450
stateProvince	X	0.672	92.594589	is_	4,780,674	43.67%	no_	8,055
county	X	3.682	22.211235	is_	6,975,852	63.72%	no_	11,063
municipality		0.008	10.96424	is_	10,409,285	95.08%	no_	17,303
locality	X	0.006	2.1544214	is_	4,968,186	45.38%	no_	353,130
verbatimLocality		74.91	93.117818	is_	8,234,354	75.21%	no_	49,462

Conclusions: Community Effort

We all have data problems

There is A LOT of variation in how we map our data

We are underutilizing what is there

Definitely new terms needed

Some easy fixes/cleanup - documentation with the recommended values and formats

Citations

Wieczorek J, Bloom D, Guralnick R, Blum S, Döring M, Giovanni R, et al. (2012) Darwin Core: An Evolving Community-Developed Biodiversity Data Standard. PLoS ONE 7(1): e29715. https://doi.org/10.1371/journal.pone.0029715

GBIF.org (09 August 2018) GBIF Occurrence Download https://doi.org/10.15468/dl.r3tqg5

GBIF.org (18 July 2018) GBIF Occurrence Download https://doi.org/10.15468/dl.k61spa

Ask me if you'd like the full citation list for all datasets!

Smithsonian Institution

NATIONAL MUSEUM of NATURAL HISTORY

Thank you

Luis Villanueva, Smithsonian Digitization Program Office

Adam Mansur, Smithsonian NMNH Department of Mineral Sciences

Stack Overflow

GBIF

All of the data providers!

EPICC collections network working on an Initial Paleo Guideline