Data Standards and how they are used for Georeferencing

Erica Krimmel • Holly Little
Talia Karim • Carrie Levitt-Bussian • Deb Paul
Local: NMNH Paleo

669,855(+-) Fossil specimen Occurrences

1970-present: Creation of digital records

Global

11,665,493 GBIF Fossil specimen Occurrences

2009/2012: Darwin Core (DwC)

❖ 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?
❖ Are new terms needed in the standard?
What is a Data Standard

- Common language
- Addresses both format and meaning
- Facilitates finding, sharing, and management of information across systems and users

Current date and time expressed according to ISO 8601

<table>
<thead>
<tr>
<th>Date</th>
<th>2020-04-27</th>
</tr>
</thead>
<tbody>
<tr>
<td>Date and time in UTC</td>
<td>2020-04-27T15:22:36+00:00</td>
</tr>
<tr>
<td>Date and time</td>
<td>2020-04-27T15:22:36Z</td>
</tr>
<tr>
<td></td>
<td>20200427T152236Z</td>
</tr>
<tr>
<td>Week</td>
<td>2020-W18</td>
</tr>
<tr>
<td>Week with</td>
<td>2020-W18.1</td>
</tr>
</tbody>
</table>

ISO 3166 - Country Codes
United States of America
Alpha-2 code: US
Alpha-3 code: USA
Biodiversity Data Standards (a sampling)

**DwC**: Darwin Core

**ABCD**: Access to Biological Collection Data

**AC**: Audubon Core

**DC**: Dublin Core

**EML**: Ecological Metadata Language

**CD**: Collection Descriptions

**GGBN Extensions**: Global Genome Biodiversity Network

**TDWG Mission**: Develop, adopt and promote standards and guidelines for the recording and exchange of data about organisms [through community effort]

https://xkcd.com/927/
**Standards vs. Guidelines**

**Standards:**
- Based on a consensus (both)
- Requirements driven
- Fixed entity with defined revision process

**Guidelines:**
- Based on a consensus (both)
- Experience driven
- Implementation of a Standard
- Evolves over time with more flexibility
Application of standards at many levels

- Global  **Workshop Day 1**
  - Ratified standard (e.g. Darwin Core, ABCD)

- Specific domain, community, or context  **Workshop Day 2**
  - Guidelines for implementation within the Paleo Community

- Local  **Workshop Day 1**
  - Mobilization - Mapping of local data
  - Standardization within institution’s CMS  **Workshop Day 2**
    - Data entry protocols
    - Locally defined Controlled vocabularies
Darwin Core (DwC)

DwC is a glossary of terms intended to facilitate the sharing of information about biological diversity by providing identifiers, labels, and definitions.

- Ratified in 2009
- Minimal scope
- Flexible structure (flat)
- 180 terms divided into 11 classes

<table>
<thead>
<tr>
<th>georeferencedBy</th>
<th>Property</th>
</tr>
</thead>
<tbody>
<tr>
<td>Identifier</td>
<td><a href="http://rs.tdwg.org/dwc/terms/georeferencedBy">http://rs.tdwg.org/dwc/terms/georeferencedBy</a></td>
</tr>
<tr>
<td>Definition</td>
<td>A list (concatenated and separated) of names of people, groups, or organizations who determined the georeference (spatial representation) for the Location.</td>
</tr>
<tr>
<td>Comments</td>
<td>Recommended best practice is to separate the values in a list with space vertical bar space (</td>
</tr>
<tr>
<td>Examples</td>
<td>Brad Millen (ROM), Kristina Yamamoto</td>
</tr>
</tbody>
</table>

https://www.tdwg.org/standards/dwc/
DwC Terms

1. **Record-Level Terms** (19 terms)
   - ex: institutionCode  basisOfRecord

2. **Occurrence** (19 terms)
   - ex: occurrenceID  recordNumber  preparations

3. **Organism** (7 terms)
   - ex: organismName  associatedOccurrences

4. **MaterialSample** (1 term)
   - ex: materialSampleID

5. **Event** (15 terms)
   - ex: fieldNumber  eventDate  samplingProtocol

6. **Location** (44 terms)
   - ex: higherGeography  locality  georeferenceRemarks

7. **GeologicalContext** (18 terms)
   - ex: group  latestAgeOrHighestStage

8. **Identification** (8 terms)
   - ex: typeStatus  identifiedBy

9. **Taxon** (33 terms)
   - ex: scientificName  class  taxonRemarks

10. **MeasurementOrFact** (9 terms)
    - ex: measurementType

11. **ResourceRelationship** (7 terms)
    - ex: relationshipOfResource

[Link to DwC Terms](http://rs.tdwg.org/dwc/terms/)
DwC and Paleo Data

- **basisOfRecord** = FossilSpecimen
- **informationWithheld** dataGeneralizations **taxonRank**
- Some data we cannot map to DwC

### GeologicalContext

<p>| geologicalContextID | earliestEonOrLowestEonothem | latestEonOrHighestEonothem | earliestEoE or LowestEoE | latestEOE or HighestEOE | earliestEO or LowestEO | latestEO or HighestEO | earliestPeriodOrLowestPeriod | latestPeriodOrHighestSystem | earliestEpochOrLowestSeries | latestEpochOrHighestSeries | earliestAgeOrLowestStage | latestAgeOrHighestStage | lowestBiostratigraphicZone | highestBiostratigraphicZone | lithostratigraphicTerms | group | formation | member | bed |
|---------------------|-----------------------------|-----------------------------|---------------------------|--------------------------|-----------------------|-----------------------|--------------------------|----------------------------|----------------------------|--------------------------|---------------------------|--------------------------|--------------------------|-----------------------------|-----------------------------|-----------------------|-----------|----------|--------|------|</p>
<table>
<thead>
<tr>
<th>DwC Term</th>
<th>NMNH EMu Field or Source</th>
<th>NMNH Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>basisOfRecord</td>
<td>added at DwC-A</td>
<td>fossilSpecimen</td>
</tr>
<tr>
<td>institutionCode</td>
<td>Museum Acronym</td>
<td>USNM</td>
</tr>
<tr>
<td>collectionCode</td>
<td>Department</td>
<td>Paleobiology</td>
</tr>
<tr>
<td>occurrenceID</td>
<td>GUID</td>
<td><a href="http://n2t.net/ar...8e">http://n2t.net/ar...8e</a> d9-f36ac0a5e423</td>
</tr>
<tr>
<td>catalogNumber</td>
<td>Prefix + CatNumber</td>
<td>PAL531308</td>
</tr>
<tr>
<td>eventDate</td>
<td>Date Visited (From)</td>
<td>2003-12-07</td>
</tr>
<tr>
<td>formation</td>
<td>formation</td>
<td>Bird Spring Fm</td>
</tr>
<tr>
<td>verbatimCoordinates</td>
<td>Original Coordinate System</td>
<td>Degrees Minutes Seconds</td>
</tr>
</tbody>
</table>
ABCD (Access to Biological Collection Data)

The ABCD Schema is an evolving comprehensive standard for the access to and exchange of data about specimens and observations (a.k.a. primary biodiversity data).

- Ratified in 2005
- Comprehensive scope
- Highly Structured (relational, hierarchical)
- GBIF: ~66 datasets within basisOfRecord = fossilSpecimen
- 780 data elements with 240 additional attributes

PaleontologicalUnit/Preservation
PaleontologicalUnit/Preservation/Completeness
PaleontologicalUnit/Preservation/Form
PaleontologicalUnit/Preservation/Matrix
PaleontologicalUnit/Preservation/Mineralization
PaleontologicalUnit/Preservation/Taphonomy
PaleontologicalUnit/TimeRange

https://abcd.tdwg.org/
The ABCD Extension EFG is a set of terms designed to represent earth science collection items (e.g. paleontological, mineralogical, and petrological).

- Developed in 2006
- Adds 840 additional terms
- Added data structures for:
  - Stratigraphy
  - Chemical analyses
  - Host rock composition
  - ...

Palaeontological_UnitType
  - PartOfOrganism
  - Completeness
  - Articulation
  - Feeding_Predation_Traces
  - Bioerosion
  - Encrustation
  - Orientation
  - Assemblage_Origin
  - Post_Burial_Transportation
  - Preservation_Quality
  - Preservation_Mode
  - Preservation_Alternation
  → (more terms)
  → (more terms)

https://github.com/tdwg/abcd/tree/master/efg
Standards for Georeference Data

USGS Localities provided by Casey McKinney, USGS
ABCD Coordinates terms

All within a container under Gathering/CoordinateSets

CoordinateSet -&gt; CoordinatesGrid -&gt; 3 more terms
CoordinateSetsUTM -&gt; 8 more terms
CoordinatesLatLong
EPISGID
Method
Notes
References
VerificationStatus
CoordinatesGridCellSystem

LongitudeDecimal  LatitudeDecimal
VerbatimLongitude  VerbatimLatitude
SpatialDatum  Accuracy
ErrorDistanceInMeters  ErrorMethod
PointRaiusSpatialFit  CoordinatesText

27 nested terms for Georeferenced Data
DwC Location Class Terms

locationID  higherGeographyID  higherGeography  continent  waterBody  islandGroup
Island  country  countryCode  stateProvince  county  municipality  locality
verbatimLocality  minimumElevationInMeters  maximumElevationInMeters
verbatimElevation  minimumDepthInMeters  maximumDepthInMeters  verbatimDepth
minimumDistanceAboveSurfaceInMeters  maximumDistanceAboveSurfaceInMeters
locationAccordingTo  locationRemarks  decimalLatitude  decimalLongitude  geodeticDatum
coordinateUncertaintyInMeters  coordinatePrecision  pointRadiusSpatialFit
verbatimCoordinates  verbatimLatitude  verbatimLongitude
verbatimCoordinateSystem  verbatimSRS  footprintWKT  footprintSRS  footprintSpatialFit
georeferencedBy  georeferencedDate  georeferenceProtocol  georeferenceSources
georeferenceVerificationStatus  georeferenceRemarks

20 terms for georeferenced data
DwC Point terms

<table>
<thead>
<tr>
<th>Point Type</th>
<th>Point Data</th>
<th>Described by</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lat/Long</td>
<td>-41.0983423, -121.1761111</td>
<td>geodeticDatum, coordinateUncertaintyInMeters,</td>
</tr>
<tr>
<td></td>
<td></td>
<td>coordinatePrecision, pointRadiusSpatialFit</td>
</tr>
<tr>
<td>UTM DMS etc..</td>
<td>41 05 54S 121 05 34W, 17T 630000 4833400</td>
<td>verbatimCoordinates, verbatimCoordinateSystem,</td>
</tr>
<tr>
<td></td>
<td></td>
<td>verbatimSRS, footprintSRS</td>
</tr>
<tr>
<td>Polygon</td>
<td>((10 20, 11 20, 11 21, 10 21, 10 20))</td>
<td>footprintWKT, footprintSpatialFit</td>
</tr>
</tbody>
</table>
DwC Georeference documentation terms

<table>
<thead>
<tr>
<th>Term</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>georeferencedBy</td>
<td>Holly Little</td>
</tr>
<tr>
<td>georeferencedDate</td>
<td>2020-04-28</td>
</tr>
<tr>
<td>georeferenceProtocol</td>
<td>Georeferencing Quick Reference Guide</td>
</tr>
<tr>
<td>georeferenceSources</td>
<td>GeoLocate</td>
</tr>
<tr>
<td>georeferenceVerificationStatus</td>
<td>verified by curator</td>
</tr>
<tr>
<td>georeferenceRemarks</td>
<td>Assumed x</td>
</tr>
</tbody>
</table>

Breadcrumbs enabling other to follow your georeferencing path. Figure out what you did to generate your results...

https://dwc.tdwg.org/terms/#dwc:georeferencedBy
YOU GET DATA! AND YOU GET DATA!
AND YOU GET DATA!

EVERYBODY GETS DATA!