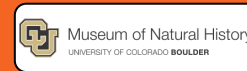

Improving the adoption and evolution of data standards for fossil specimens

Talia Karim, Erica Krimmel, Holly Little

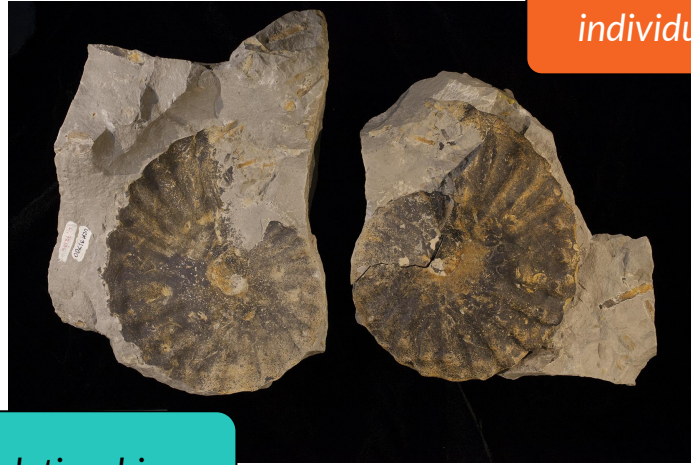
CC-BY 4.0 | <https://doi.org/10.3897/biss.5.75646>



organismQuantity



individualCount



ResourceRelationship

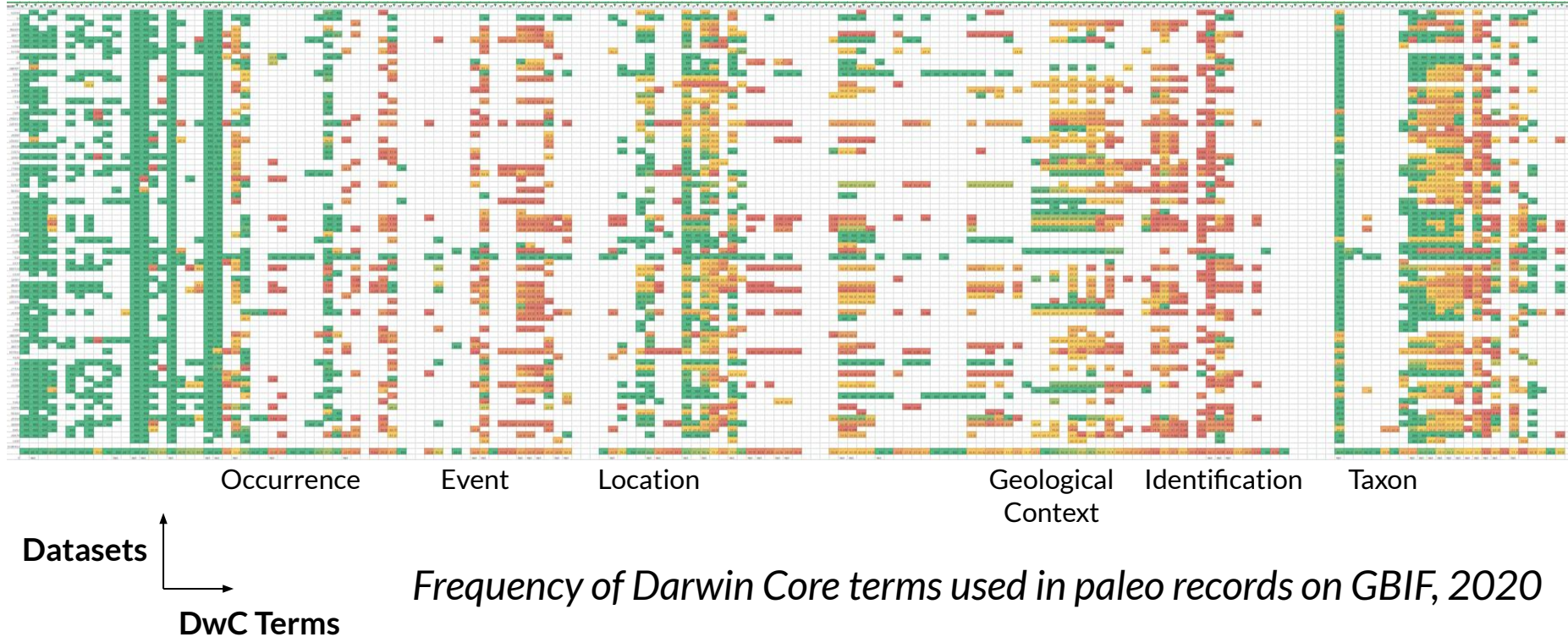
Taxon



Location



Visualizing Implementation of DwC



Group 1

Challenge: Terms are clear to implement, but underutilized or inconsistently used

Solution: Better education at the data provider-level

EXAMPLE

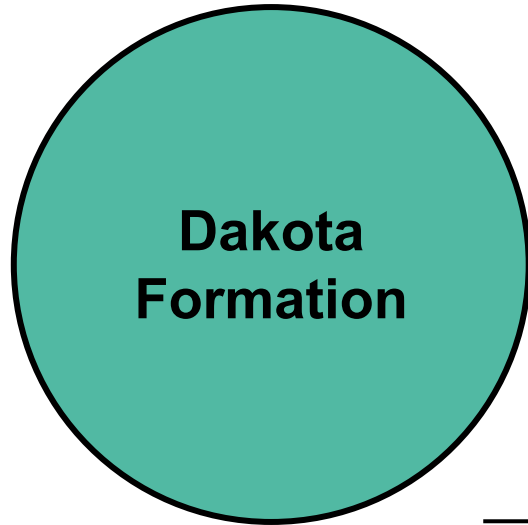
1.1 million fossil specimen records in GBIF record lat/long with a maximum precision of 0.01, but only ~5% of these records have any value present in *dwc:informationWithheld* or *dwc:dataGeneralizations*.

Guidelines for Sensitive Localities

DwC Term	Paleo Guideline	Paleo Example
decimalLatitude decimalLongitude	May be truncated for paleontological specimens. If an institution truncates these values, they should also serve dataGeneralizations with an explanation such as: “Latitude and longitude reported at maximum precision of 0.1 degrees.”	
informationWithheld	This is an important field to include for paleontological specimens. In many cases, specific locality information will be restricted for some or all paleontological specimens due to federal regulations as well as the preferences of private landowners.	Example: “More data may be available”
dataGeneralizations	This is an important field to include if an institution does not serve the most specific decimal latitude/longitude available for a specimen. It is common to redact or fuzz geographic information to protect fossil localities from theft.	Example: “Latitude and longitude reported at maximum precision of 0.1 degrees.”

Standard Formatting & Controlled Vocabularies

e.g. to use with terms in the GeologicalContext class



Group 2

Challenge: Terms may be clear to implement, but terminology used to describe and define them is unfamiliar to paleontologists or read as unnecessary for fossil occurrences

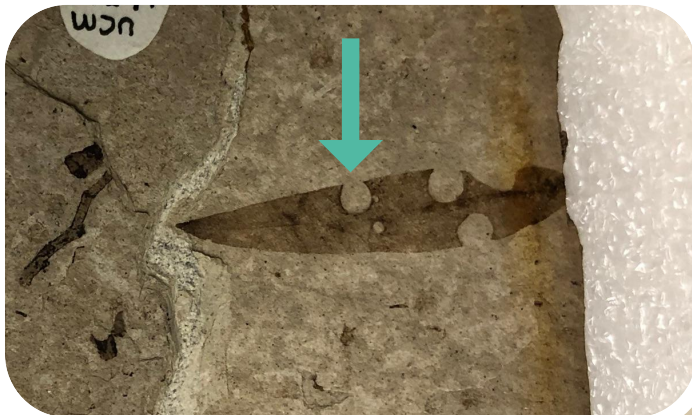
Solution: Improved term documentation that is inclusive of the paleo context

EXAMPLE

The examples for *dwc:relationshipOfResource* were primarily neontological relationships. Adding the example “on slab with” highlights that this field is relevant to paleo collections.



ResourceRelationship | associatedOccurrences



Change term - relationshipOfResource

Examples: sameAs, duplicate of, mother of, offspring of, sibling of, parasite of, host of, valid synonym of, located within, pollinator of members of taxon, pollinated specific plant, pollinated by members of taxon, on slab with

Group 3

Challenge: Terms do not adequately represent information for the paleo context

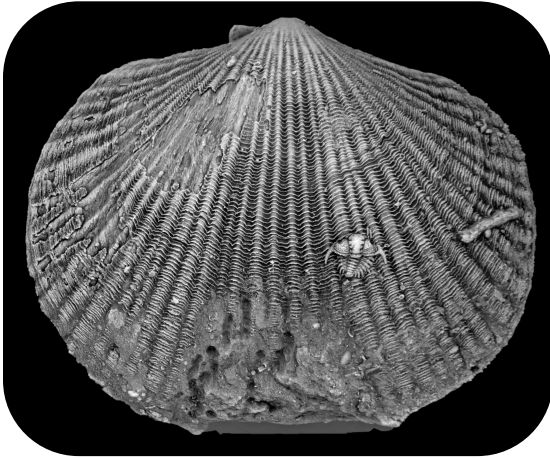
Solution: Paleo community must participate in standards development and review process

EXAMPLE

Using *dwc:individualCount* for paleo specimens is complicated by the preservation of multiple biological organism in a single physical object, and of a single biological organism in multiple physical objects.



individualCount | organismQuantity



Complex Topics Span All Groups

Challenge

Problems recording **taxonomic data** in Darwin Core are complex, with solutions that are multiple and interrelated

Group 1

Data affected when aggregators weight terms that data providers underutilize (e.g., *dwc:taxonRank*)



Group 2

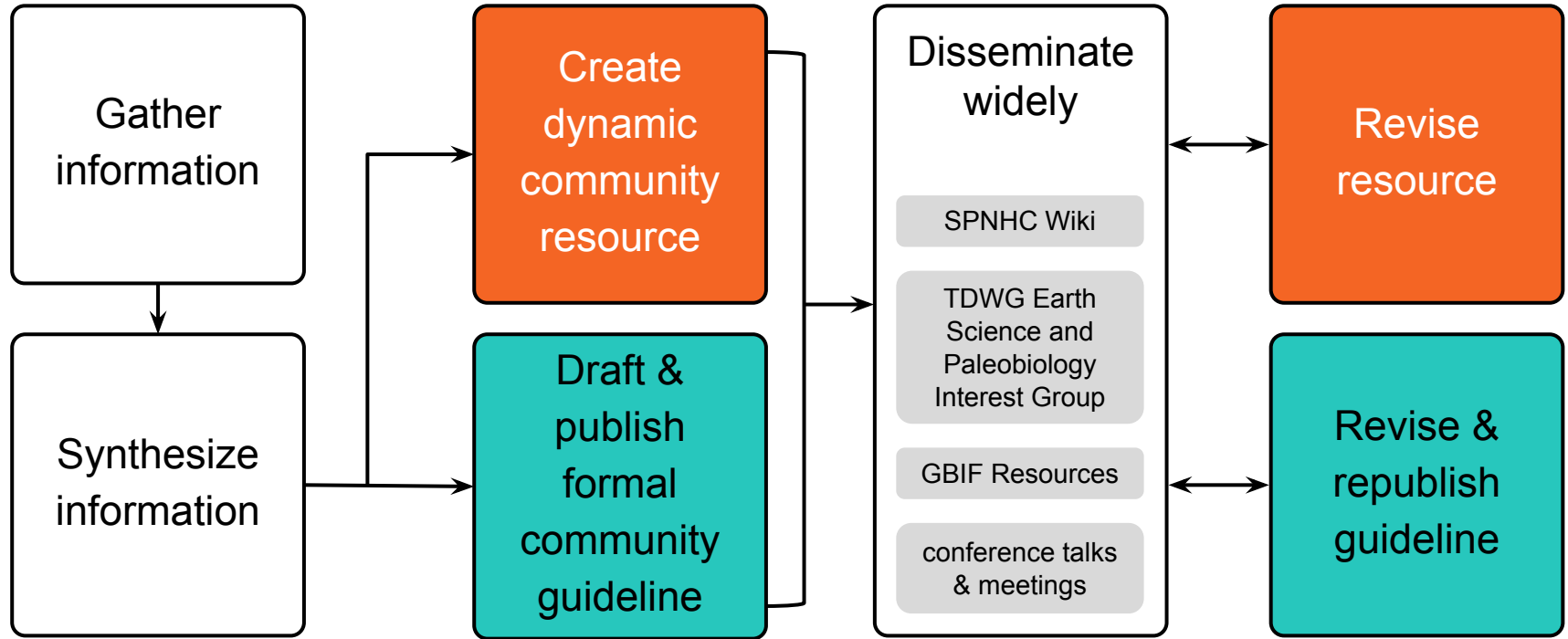
Unclear how or where to record “incertae sedis” for ichnotaxa



Group 3

Nowhere to share data about taxonomic ranks commonly used in paleo (e.g. subclass, clade)

Community Knowledge Management



Digitally Accessible Data is Essential for Research



Basis of record	Count	
Observation	19,881,036	
Machine observation	16,419,890	
Human observation	1,594,700,333	
Material sample	37,027,700	
Literature	978,964	
Preserved specimen	192,393,497	
Fossil specimen	11,899,181	
Living specimen	1,725,851	
Unknown	15,544,451	

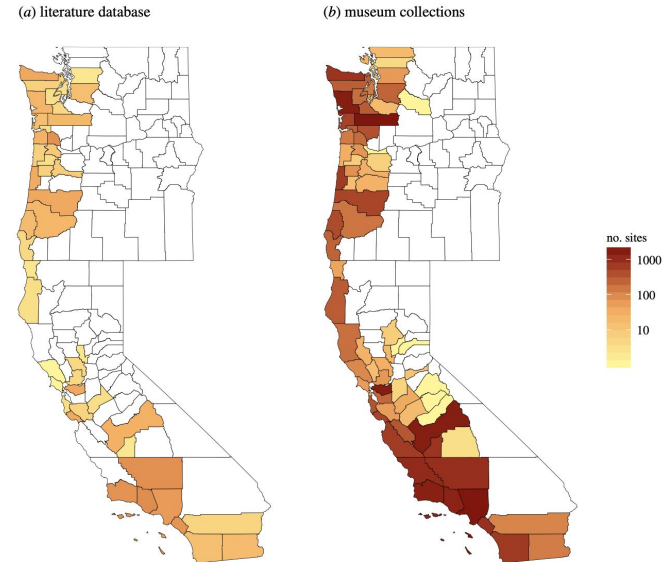
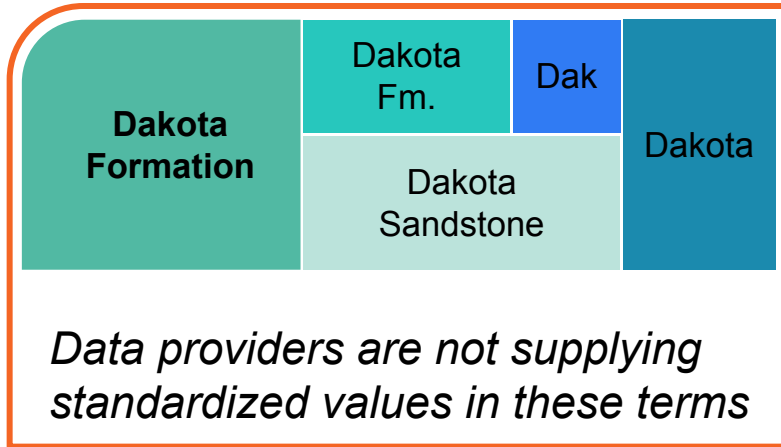


Figure 1. Visualization of the 23-fold increase in digitally accessible Cenozoic marine invertebrate palaeontological collection sites (26 059) from museum collections compared with the number of collection sites (1139) from literature data currently entered into the PBDB (<https://paleobiodb.org/>) for California, Oregon and Washington. (a) Number of sites per county currently included in the PBDB (<https://paleobiodb.org/>); (b) number of sites per county now digitally mobilized across nine institutions of the EPICCC TQN (<https://epiccc.berkeley.edu/>). The number of sites per county for each map are provided in the Supplemental_Data.csv file deposited in the Dryad data repository ([doi:10.5061/dryad.j0r8127](https://doi.org/10.5061/dryad.j0r8127)) [11].

Improving Implementation of Data Standards is Essential for Research

Researchers cannot currently discover data using the entry points they expect, e.g. litho- or chronostratigraphy



GBIF

Data aggregators are not indexing these terms to make them searchable

Thank you

Continue the discussion at the TDWG Earth Sciences and Paleobiology Interest Group Meeting - November 4th @ 17:30 UTC

Resources

Paleo Digitization Working Group Wiki, with links to more information about our group's "Happy Hour" topics and Slack workspace: <https://bit.ly/paleo-digi>

Krimmel E, Karim T, Little H, Walker LJ, Burkhalter R, Byrd C, Millhouse A, Utrup J (2021) The Paleo Data Working Group: A model for developing and sustaining a community of practice. *Biodiversity Information Science and Standards* 5: e74370. <https://doi.org/10.3897/biss.5.74370>

Little H (2018) Establishing a New Framework for Paleontological Data Through an Evaluation of Current Data Sharing Practices. *Biodiversity Information Science and Standards* 2: e25437. <https://doi.org/10.3897/biss.2.25437>

Wieczorek J, Bloom D, Guralnick R, Blum S, Döring M, 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>
