

Improving the Adoption and Evolution of Data Standards for Fossil Specimens

Holly Little[‡], Talia S. Karim[§], Erica Kimmel^{||}

[‡] Smithsonian National Museum of Natural History, Washington, DC, United States of America

[§] University of Colorado, Boulder, United States of America

^{||} iDigBio, Florida State University, Tallahassee, United States of America

Corresponding author: Holly Little (littleh@si.edu)

Abstract

As we atomize and expand the digital representation of specimen information through data standards, it is critical to evaluate the implementation of these developments, including how well they serve discipline-specific needs. In particular, fossil specimens often present challenges because they require information to be captured that is seemingly parallel to, but not entirely aligned with, that of their extant counterparts. Previous work to evaluate data sharing practices of paleontology collections has shown an imbalance in the use of Darwin Core (DwC) (Wieczorek et al. 2012) terms and many instances of underutilized terms (Little 2018). To expand upon that broad assessment and encourage better adoption of evolving standards and data practices by fossil collections, a more in-depth review of term usage is necessary. Here we review specific DwC terms that are underutilized or that present challenges for fossil occurrence records, and we examine the subsequent impact on data discovery of paleo specimens. We conclude by sharing options for improving standards implementation within a paleo context.

We see key patterns and challenges in current implementation of DwC in paleo collections, as evidenced by evaluations of the typical mappings found in occurrence records for fossil specimens, data flags applied by aggregators, and discussions within the paleo collections community. These can be organized into three broad groupings.

Group 1: Some DwC terms (or classes of terms) are clear to implement, but are underutilized due to issues that are also found within the neontological community. Example: Location. In the case of terms related to the Location class, paleontology has a need for a way to deal with sensitive locality information. The sensitivity here typically relates to laws restricting the sharing of locality information to protect fossil sites versus neontological requirements to protect threatened, rare, or endangered species. The end goal of needing to fuzz locality information without completely making the specimen record undiscoverable or unusable is the same. There is a need for better education at the paleo data provider-level related to standards for recording and sharing information in this category, which could be based on existing neontological community standards.

Group 2: A second group of DwC terms often seem clear to implement, but the terminology used to describe and define them might be unfamiliar to paleontologists or read as unnecessary for fossil occurrences. This uncertainty about the applicability of a term to paleo data can often result in data not being mapped or fully shared. Example: recordedBy (= collector). In these cases, a simple translation of what the definition means in verbiage that is familiar to paleontologists, or the inclusion of paleo-oriented examples in the DwC documentation, can make implementation clear.

Group 3: A third group of issues relates to DwC terms, classes, and/or extensions that are more complicated in the context of fossil vs. neontological data. In some cases use of these terms is complicated for neontological data as well, but perhaps for different reasons. The terms impacted by these challenges can sometimes have the same general use, but due to the nature of fossil preservation, or because a term has a different meaning within the discipline of paleontology, additional layers of uncertainty or ambiguity are present. Examples: Resource Relationship/Interactions, Individual count, Preparations, Taxon. Review of these terms and their related classes and/or the extensions they are part of has revealed that they might require qualification, further explanation, additional vocabulary terms, or even the need for special handling instructions when data are ingested and normalized at the aggregator level. This group of issues is more complicated to resolve, but the problems are not intractable and can progress toward solutions through further discussion within the community, active participation in the standards development and review process, and development of clear guidelines.

Strategically assessing these terms and generating discipline-specific guidelines to be used by the paleo community can improve the mobilization and discovery of fossil occurrence data. Documenting these paleo data practices not only helps data providers, it also increases the utility of these data within the broader research community by clearly outlining how the terms were used. Overall, this discipline-focused approach to understanding the implementation of data standards like DwC at the term level, helps to increase knowledge sharing across the paleo community, improves data quality and standards adoption, and moves these datasets towards alignment with best practices like the [FAIR](#) (Findable, Accessible, Interoperable, Reusable) data principles.

Keywords

paleontology, fossil occurrences, Darwin Core, community guidelines, data mobilization, implementation

Presenting author

Holly Little, Talia Karim, Erica Krimmel

Presented at

TDWG 2021

Conflicts of interest

References

- Little H (2018) Establishing a New Framework for Paleontological Data Through an Evaluation of Current Data Sharing Practices. Biodiversity Information Science and Standards 2 <https://doi.org/10.3897/biss.2.25437>
- Wieczorek J, Bloom D, Guralnick R, Blum S, Döring M, Giovanni R, Robertson T, Vieglais D (2012) Darwin Core: An Evolving Community-Developed Biodiversity Data Standard. PLOS One 7 (1). <https://doi.org/10.1371/journal.pone.0029715>