What Matters for an occurrenceID and What Is an occurrenceID That Matters?

Yi-Ming Gan[‡], Abigail Benson[§], Emilio Mayorga[|], Jonathan Derek Pye[¶], Stephen Formel[#]

‡ Royal Belgian Institute of Natural Sciences, Brussels, Belgium
§ U.S. Global Change Research Program, Menifee, United States of America
| University of Washington, Seattle, United States of America
¶ Ocean Tracking Network, Halifax, Canada
# U.S. Geological Survey, New Orleans, United States of America

Corresponding author: Yi-Ming Gan (ymgan@naturalsciences.be)

Abstract

In the Darwin Core data standard (Darwin Core Maintenance Group 2023), the concept of dwc:Occurrence (Wieczorek et al. 2012) in ecological data presents a data model construct not commonly found in data management practices used by ecological data collectors. We frequently encounter raw data without an Occurrence table. For example, the concept of Occurrence can be represented as a cell, where the rows represent sampling sites, the columns represent species, and the value of each cell indicates the count of the species at a specific site. A value in a cell in such a matrix can be interpreted as x number of individuals of species y occurred at sampling site z.

While data providers tend to track data on tangible individual components (e.g., species, location, sample), generating "Occurrence records" typically requires pivoting and/or joining tables associated to these components. Maintaining a stable and persistent occurrenceID for an Occurrence record created through data transformation is not an easy task. This is especially true for long-term monitoring datasets, where the underlying tables used to generate Occurrence records are continuously updated.

Additionally, most ecological data collectors are focused on the primary use of the data, not on the long term integration and accessibility of the data. The Occurrence concept is only required in data exchange format but not needed in ecological data management practices. The disconnect between the practical data management needs of data collectors and the abstractions required for data exchange raises challenges, particularly with an increasing expectation for globally unique and persistent occurrenceIDs.

This presentation will explore the difficulties of creating and managing occurrenceIDs for data providers and managers, especially those who manage data using basic systems such as spreadsheets and simple relational databases. Maintaining stability and persistence of identifiers for inherently artificial constructs like Occurrences within the original, component-based data structure can pose significant challenges. We will

explore why meaningful identifiers for occurrenceIDs are often preferred by data providers. We will unpack different use cases and delve into how and why occurrenceIDs were constructed for each use case. Through this discussion, we hope to spark a conversation that informs future data modeling efforts and addresses the inherent artificiality of Occurrences.

Keywords

biodiversity data standard, Darwin Core, occurrence, data modeling, identifier

Presenting author

Yi-Ming Gan

Presented at

SPNHC-TDWG 2024

Acknowledgements

We would like to thank various members of the Standardizing Marine Biological Data community who participated in the discussions around the topic of occurrenceID as outlined in this abstract. We appreciate the input from the community, including the co-authors, as well as Margaret O'Brien and Dean Pentcheff.

This abstract was improved for flow and grammar using ChatGPT and Gemini.

Conflicts of interest

The authors have declared that no competing interests exist.

References

- Darwin Core Maintenance Group (2023) Darwin Core List of Terms. Biodiversity Information Standards (TDWG). URL: http://rs.tdwg.org/dwc/doc/list/2023-09-18
- 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