Biodiversity Information Science and Standards 8: e141876
doi: 10.3897/biss.8.141876

Conference Abstract

Use Cases Help to Identify Primary Concepts in Biodiversity Information Modeling

Stanley Blum ‡

‡ Biodiversity Information Standards (TDWG), San Francisco, United States of America

Corresponding author: Stanley Blum (stanblum@gmail.com)

Received: 15 Nov 2024 | Published: 15 Nov 2024

Citation: Blum S (2024) Use Cases Help to Identify Primary Concepts in Biodiversity Information Modeling. Biodiversity Information Science and Standards 8: e141876. https://doi.org/10.3897/biss.8.141876

Abstract

“Use cases” are “a methodology used in system analysis to identify, clarify and organize system requirements” (Brush 2022). They provide context and purpose for data concepts. The Darwin Core (Wieczorek et al. 2012) was initially designed to serve two purposes: 1) to gather data for documenting species distributions, particularly through species distribution modeling, and 2) to support the discovery of specimens in biological collections. The occurrence concept, the "existence of a dwc:Organism at a particular place at a particular time" (Darwin Core Maintenance Group 2009), was established so that both observations and specimens could be combined into a single tabular data set and used to document the distribution of a species.

By 2008, the TDWG Technical Architecture Group began recommending that TDWG develop its standards in the framework of the semantic web (SW); i.e., RDF. Two efforts have contributed significantly to casting the terms of Darwin Core in RDF. Baskauf and Webb (2015) distilled nearly a decade of online discussions into the Darwin-SW. In a separate effort, but involving many of the same people, a series of workshops were convened to coordinate standards between the genomics and biodiversity communities and produced the BioCollections Ontology (BCO; Walls et al. 2014), which placed Darwin Core terms in the context of the "OBO foundry" (Smith et al. 2007) and the Basic Formal Ontology (BFO; Arp et al. 2015, but also see videos by Smith).

Darwin-SW included explicit recognition of IndividualOrganism, Occurrence, Event, and Token (i.e., evidence of a dwc:MaterialSample or an observation). The simplest use case,

© Blum S. This is an open access article distributed under the terms of the Creative Commons Attribution License (CC BY 4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

in which an organism is collected or observed once, doesn't require that Organism, Occurrence and MaterialSample/Observation be recognized as separate entities. The relationships are one-to-one-to-one. They can be joined into a single entity with only one identifier (e.g., materialSampleID) without loss of information. Note that a specimen or observation infers the existence of an organism and its occurrence in nature. The use cases that require separating Organism and possibly Occurrence from the MaterialSample or Observation are the ones where an organism is sampled or observed, remains in nature, and is subsequently sampled or observed again; i.e., the organism is the target of more than one dwc:Event. These cases require that the organism can be reliably identified as the same organism encountered earlier; e.g., by tag, identifying marking, DNA fingerprint, or precise and fixed location for sessile organisms.

The BCO (Walls et al. 2014) does not explicitly recognize something equivalent to the dwc:Occurrence class. Following Smith et al. (2007), the focus is on "realism," things and relationships in the real world, as opposed to what we want our information systems to do. At a high level, the BFO separates material things, processes, and information artifacts as fundamentally different entities. Accordingly, the BCO separates the "material sampling process" from the "observing process," as observations and material samples are fundamentally different. Again, an occurrence class is not present because it represents a union of MaterialSample and Observation; it is justified by an analytical use-case, which was not in scope. That might not be ultimately disqualifying, but it raises a caution flag.

A question then emerges for original providers who practice only the simplest case: should the provider manufacture dwc:occurrenceID and dwc:organismID even if they aren’t used in the original database? If they are useful to someone outside the local context, should creating redundant identifiers be the responsibility of the provider or the aggregator?

Fig. 1A shows how the concepts in Darwin-SW could be represented in an entity-relationship diagram, with the Occurrence entity used to realize the many-to-many relationship between Organism and Event. Fig. 1B shows an alternative model, in which MaterialSample realizes the relationship between Organism and Event (the Observation entity is not shown, but would parallel MaterialSample, realizing another association between Organism and Event). If there are no attributes that are most appropriately assigned to Occurrence, this representation could be viewed as simpler and sufficient.

The Darwin Core Quick Reference Guide lists 25 properties of the Occurrence class. My contention is that all but a few would be more appropriately assigned to the MaterialSample or Observation. Note that the MaterialSample or Observation represents the Organism at the time of the Event, and can be viewed as the appropriate subject for properties that change over time, e.g., lifeStage and reproductiveCondition. Moreover, others have argued that even permanent features of an Organsim are more correctly represented as having been directly assessed in the MaterialSample/Observation. It allows for contradictory assessments, but accommodating and resolving contradictions

are real parts of scientific research. The alternative placements of properties not assigned to MaterialSample/Observation are:

- occurrenceID: deprecated;
- recordedBy and recordedByID: move to Event
- georeferenceVerificationStatus: move to Location or Event.

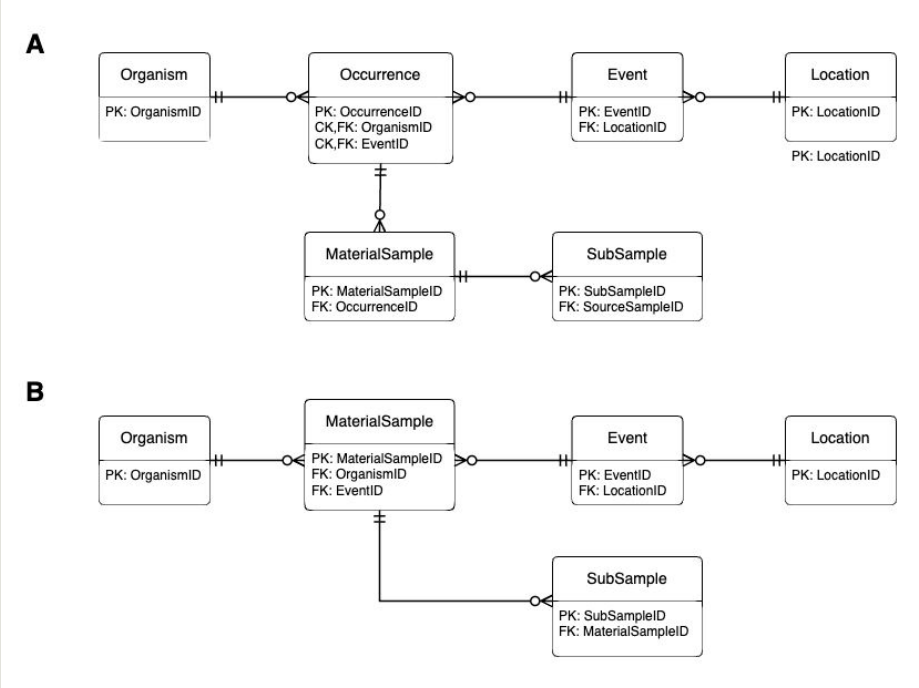

Figure 1.

Alternative entity relationship diagrams representing the concepts in Darwin-SW under discussion here. Primary (PK) and foreign (FK) keys are listed in entities. A) With "Occurrence" represented as the association entity between Organism and Event, producing a MaterialSample. B) With the MaterialSample entity establishing or embodying the relationship between Organism and Event.

Under the model represented (in part) by Fig. 1B, the task of forming the union between MaterialSample and Observation (for documenting species distributions) would fall to the aggregator or end user. The important point is that where these unions are created in our biodiversity pipelines is an engineering choice. The dwc:Occurrence class is a term of convenience, not necessarily a reflection of real-world things and processes.

Keywords

Darwin Core, occurrence, ontology, Semantic Web

Presenting author

Stanley Blum

Presented at

SPNHC-TDWG 2024

Conflicts of interest

The authors have declared that no competing interests exist.

References

- Arp R, Smith B, Spear A (2015) Building Ontologies with Basic Formal Ontology. The MIT Press, 248 pp. [ISBN 9780262527811]
- Baskauf S, Webb C (2015) Darwin-SW: Darwin Core-based terms for expressing biodiversity data as RDF. https://www.semantic-web-journal.net/system/files/swj1213.pdf. Accessed on: 2024-10-21.
- Brush K (2022) Use Case. https://www.techtarget.com/searchsoftwarequality/definition/use-case. Accessed on: 2024-10-17.
- Darwin Core Maintenance Group (2009) Darwin Core Quick Reference Guide. https://dwc.tdwg.org/terms/#occurrence. Accessed on: 2024-10-17.
- Smith B, Ashburner M, Rosse C, et al. (2007) The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. Nature Biotechnology 25 (11): 1251-1255. https://doi.org/10.1038/nbt1346
- Walls R, Deck J, Guralnick R, et al. (2014) Semantics in Support of Biodiversity Knowledge Discovery: An Introduction to the Biological Collections Ontology and Related Ontologies. PLoS ONE 9 (3). https://doi.org/10.1371/journal.pone.0089606
- Wieczorek J, Bloom D, Guralnick R, et al. (2012) Darwin Core: An Evolving Community-Developed Biodiversity Data Standard. PLoS ONE 7 (1). https://doi.org/10.1371/journal.pone.0029715