Darwin Core Data Package (DwC-DP) Implementation Experience and Feature Report

Authors: Kate Ingenloff, Cecilie Svenningsen, Yi Ming Gan, John Wieczorek, Tim Robertson, Paula Zermoglio

NOTE: Because the introduction of this document provides the rationale for a vocabulary enhancement to Darwin Core and describes how its necessary features were determined, it serves as the Feature Report required by Section 4.2.1 of the TDWG Vocabulary Maintenance Specification as well as the Implementation Experience Report as required by Section 4.2.2 of the TDWG Vocabulary Maintenance Specification for that enhancement.

Table of Contents

Introduction & background

The Darwin Core Data Package (DwC-DP) originated from an effort led by the Global Biodiversity Information Facility (GBIF) to enhance its scientific relevance by expanding its underlying data model. The effort to develop an improved data model involved exploring a series of data use cases, including biological survey and monitoring data, eDNA and metabarcoding data, camera trap data, data associated with tissue samples, specimen data with media, citizen science data from iNaturalist, automatic moth trap data, and organismal interactions data. The first product of that endeavor was a conceptual "Unified Model", which simultaneously accommodated a wide variety of documented use cases for diversification (**Figure 1**).

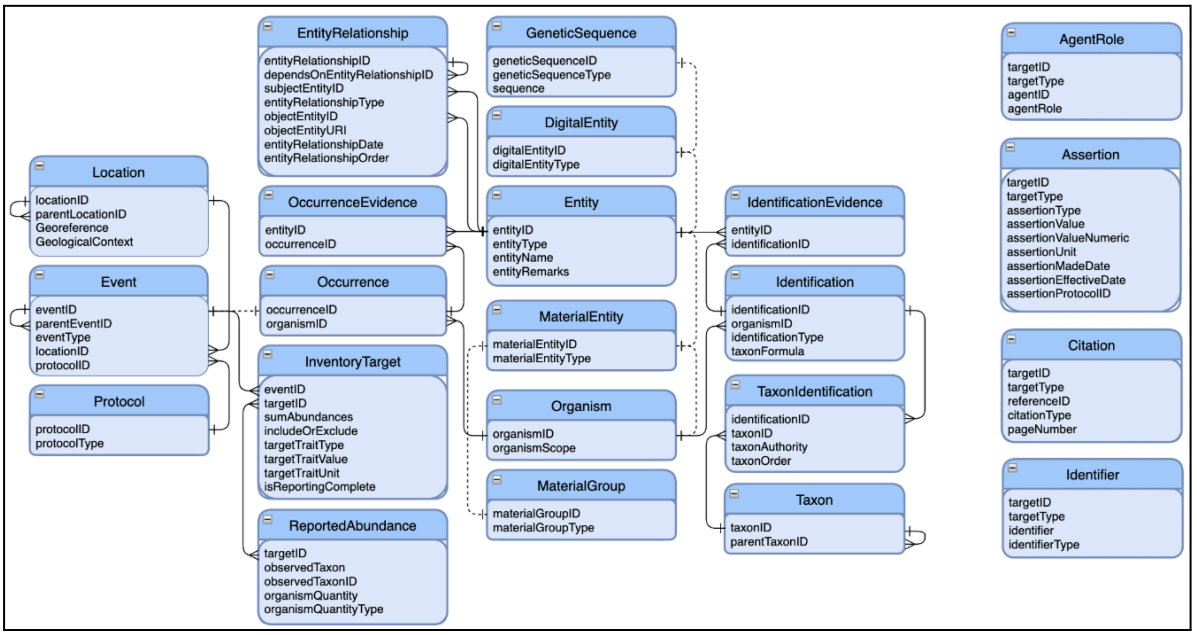

Figure 1. Schematic of the Unified Model as an entity relationship diagram.

The Unified Model was presented to the community in a series of webinars and other presentations between 2021 and 2023 and tested by mapping datasets through some targeted community efforts. While sufficiently flexible to capture each of the data types evaluated, the general consensus of the broader biological data publishing community was that the model was prohibitively complex. The community provided clear feedback that any new data publishing model should:

- be detailed enough to capture complex event-based data structures to effectively support nested survey designs and data types from across biological domains,
- be accessible to the data publishing community in that the complexity is great enough to capture as many data types as possible, but maintain a level of familiarity and consistency with the existing data publishing model, Darwin Core Archive (DwC-A),
- overcome the limitations of DwC-A, specifically the star schema,
- minimise the number of different data publishing models,
- capture thorough dataset metadata, including project-level and data provenance information,
- ensure that published data meet FAIR standards, and

- support relatively smooth translation of existing published datasets implementing DwC-A to the new data model.

In light of this feedback, the Darwin Core Conceptual Model (**Figure 2**) emerged as a practical, "just enough" simplification of the Unified Model. The Darwin Core Data Package (DwC-DP) is an implementation of the Conceptual Model using a Frictionless Data Package format. It is simple and similar enough to the existing DwC-A to support community adoption while maintaining the complexity necessary to capture a much greater breadth and depth of information.

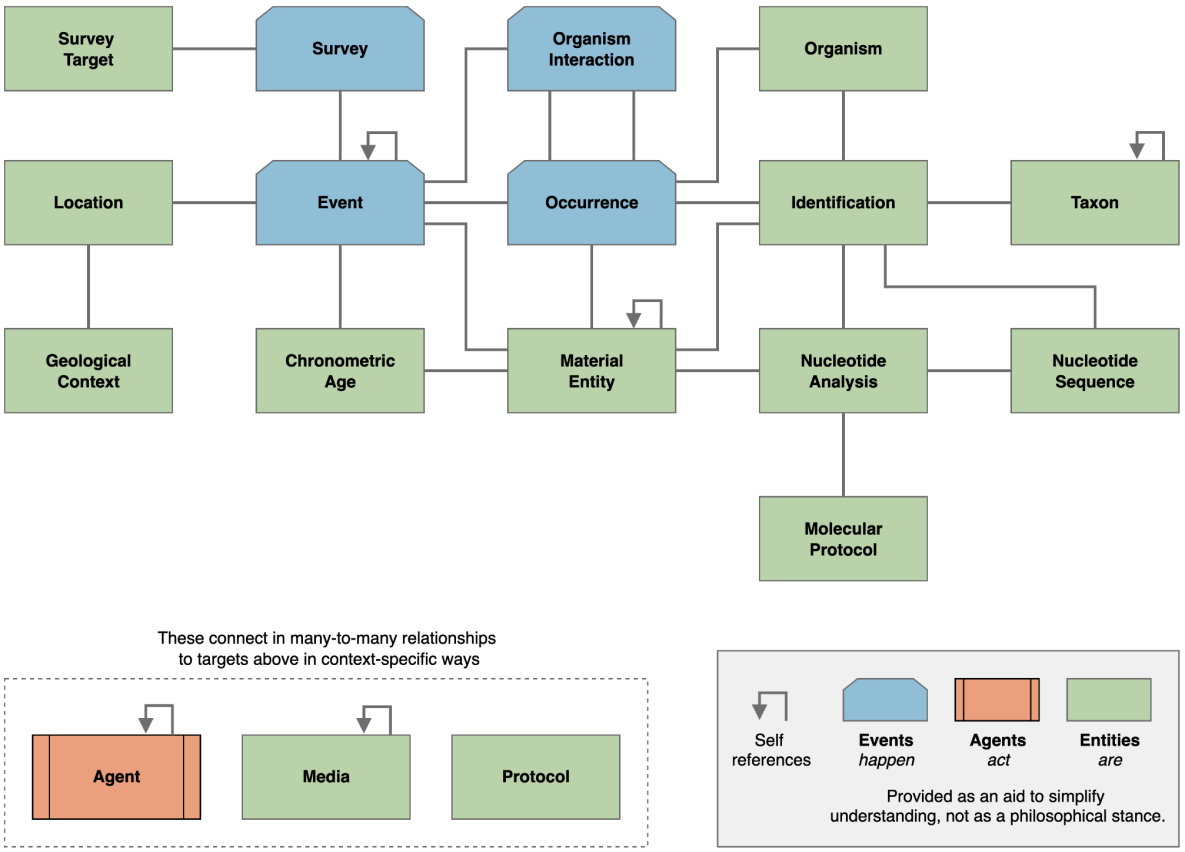

Figure 2. Schematic of the Darwin Core Conceptual Model.

DwC-DP is proposed as a vocabulary enhancement to the Darwin Core Biodiversity Data Standard. It is a standards-based format designed to support richer and more structured biodiversity data publishing than is possible with DwC-A. Like DwC-A, DwC-DP is a Darwin Core-based specification for how to structure and share biodiversity data. It consists of a JSON file that describes the structure of the Data Package, including an EML file to describe the dataset metadata, and the text files (formatted, for example, in CSV) that contain the data. Unlike DwC-A, DwC-DP does not use 'cores' to structure data, and the structural complexity limitation of the star schema is not an issue. In addition, new classes and connections to and between them are part of the conceptual model that DwC-DP enables in practice. This means that DwC-DP can support new kinds of data that DwC-A cannot, such as Organism Interactions, environmental DNA surveys, Identifications that support multiple taxa, and Surveys in which detections, non-detections and abundance can be expressed for any user-defined target scope, among many others.

Early use cases

More than 20 use cases were explored in the development of the broader conceptual model. Information about 14 of these is available under 'Case Studies' on the GBIF New Data Model page. Seven case studies were presented as part of five public webinars/presentations:

- Exploring camera trap data: This webinar presents a case study exploring camera trap data from CamtrapDP.
- Exploring collection management systems: This webinar covers three case studies exploring specimen data, including specimen data in the Arctos collection management system, specimen data with tissue samples (e.g., VertNet data), and specimen data with media.
- Exploring eDNA data: This webinar reports on a case study exploring eDNA data from soil samples reported by the BioWide project. The original BioWide dataset is available at GBIF as an occurrence dataset.
- Exploring interactions data on plant pollination: This webinar uses a *Melitaea cinxia* (Glanville fritillary butterfly) population study on Sottunga Island as a use case to report on explorations into capturing organism interaction data. The use case is based on data from Duplouy A., A. Nair, T. Nyman, and S. van Nouhuys. 2021. Long-term spatiotemporal genetic structure of an accidental parasitoid introduction, and local changes in prevalence of its associated Wolbachia symbiont. *Molecular Ecology* 30(18), 4368-4380. https://doi.org/10.1111/mec.16065.
- Invasive alien species in the GBIF Unified Model: This presentation reports on experiences in developing a use case about the management of invasive alien taxa in Flanders – muskrats (*Ondatra zibethicus*).

One case study exploring wildlife disease data from eNetWild was published as a European Food Safety Authority (EFSA) Supporting publication: ENETWILD consortium, F. Jaroszynska, G. Body, S. Pamerlon, and A.S. Archambeau. 2022. Applying the Darwin Core data standard to wildlife disease – advancements toward a new data model. *EFSA Supporting Publications* 19(11):7667E. https://doi.org/10.2903/sp.efsa.2022.EN-7667.

DwC-DP use cases

DwC-DP was tested and subsequently enhanced through a series of use cases representing multiple biological domains and data types. Many of the case studies used to develop the broader conceptual model in early efforts were re-opened to test DwC-DP; and numerous others were added. These use cases consist of original datasets that were structured (mapped) to test the Darwin Core Data Package publishing model. Once a first, stable version of DwC-DP was achieved, a webinar formally introduced DwC-DP and a preliminary community testing period was announced. This community testing period was open from the beginning of May 2025 until 1 August 2025. Parties interested in exploring the DwC-DP were provided access to a Data Mapping Guide and a GitHub repository dedicated to the early testing phase. A curated suite of examples were made available in the DwC-DP examples folder. The repository includes example datasets broadly representing three data categories: biological surveys, datasets with material samples, and datasets reporting organism

interactions. There are three example datasets for each dataset category. Four datasets were added to the DwC-DP example datasets repository (https://github.com/gbif/dwc-dp-examples) and/or published using the GBIF test instance of DwC-DP integrated publishing toolkit (IPT) during the early community review.

An in-depth overview of the DwC-DP implementation experience for two use cases used in early data model development and in testing the Darwin Core Data Package–the Broke West Fish dataset and the Insektmobilen dataset–is described below.

Broke West Fish, OBIS

'Broke West Fish' is a test dataset derived from a biotic survey with target organism scopes (fish collected by rectangular midwater trawl nets) published by the Ocean Biodiversity Information System (OBIS). This dataset has been a use case several times. It was used as a case study in early data model explorations, and it was used in developing and testing the Humboldt extension. Finally, the dataset was used to explore the implementation of the Darwin Core Data Package.

The original dataset is available here. The DwC-DP example dataset folder includes a ReadMe document, the originally mapped data used for testing against the database schema, and output data. The test dataset was successfully published to the GBIF test instance of the DwC-DP IPT and is viewable at broke-west-fish on the DwC-DP test IPT.

The Broke West test dataset was used to explore:

Inclusion of persistent identifier terms in the Survey table. Although IRI equivalents exist for terms in the Humboldt extension, the current GBIF IPT implementation does not allow publication of these identifiers. The use of identifiers was tested as a means of ensuring coherence in the shared data, since string values alone may vary (e.g., name spellings). By including persistent identifiers such as ORCID for *samplingPerformedByID* alongside the name of the person in *samplingPerformedBy* in the Survey table, proper attribution of contributors can be maintained.

Representation of survey targets. In the Humboldt extension, multiple survey targets are reported in the same table, requiring survey design and effort information to be repeated for each target combination, resulting in clutter. Moreover, survey targets can be expressed from different perspectives, such as listing all taxa within a scope or defining a functional group at the survey area with a given body size range that the sampling gear is capable of detecting. While this can be forced into the Humboldt extension using *eco:verbatimTargetScope*, the Survey Target table in DwC-DP provides a more structured and flexible approach for capturing such information.

Representation of community and individual measurements. In marine datasets, measurements are often collected both on complete groups of individuals and on individuals subsampled from a group. For example, the abundance of a fish species within a catch is recorded as a measurement of an Occurrence, while individual specimens from that catch (with their own sample IDs) are further measured for standard length, total length, wet weight, and other traits. Under DwC-A, the common workaround has been to duplicate Occurrence records to accommodate both group- and individual-level measurements (**Figure 3**). DwC-DP provides a clearer distinction between Occurrence

and Material, enabling separate identifiers and measurement assertions to be maintained without duplication.

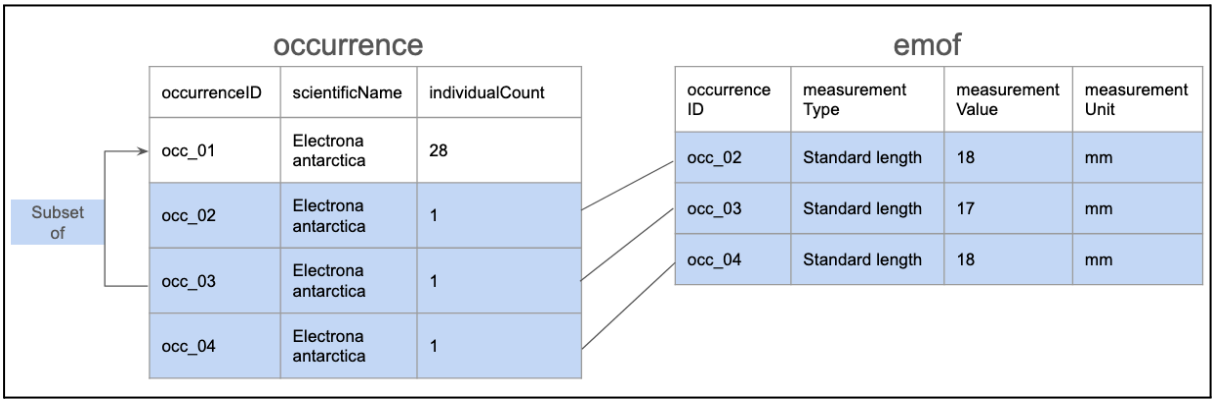

Figure 3. Representation of group- and individual-level measurements in DwC-A. A catch of *Electrona antarctica* (occ_01) is recorded with an individual count of 28, while subsets (occ_02 – occ_04) are duplicated as Occurrence records to capture individual measurements (e.g., standard length) in the extended measurement or fact (emof) table.

Information constrained by the DwC-A star schema. During the mapping process, some information was never published because it could not be easily represented in DwC-A. For instance, the stomach contents of a subset of fish were identified and linked to individual counts and digestion state (e.g., two krill observed in the stomach of specimen 001, itself an individual within an Occurrence of 25 fish). Such nested relationships are difficult to capture within the DwC-A structure. Similarly, specimen photographs could be represented in DwC-A, but the model forces them into the Occurrence core using an audiovisual or simple multimedia extension, which means sampling event information will have to be repeated for all Occurrences. Although associatedMedia can be used within the Occurrence extension alongside Event core, this results in a loss of critical attribution information. This information is not constrained in this way with DwC-DP.

Preservation of identifier meaning. In DwC-A, publishing through an IPT results in an Occurrence table that requires a unique *occurrenceID*, leading to conflation of the Occurrence and Material classes. For example, a preserved specimen (an individual fish from a catch of 25) is represented as an Occurrence record with its own *occurrenceID* and *basisOfRecord* = 'PreservedSpecimen', even though it originates from a single catch-level Occurrence. In practice, the *occurrenceID* is often repurposed to serve as the identifier of the specimen, since the table structure leaves no alternative. In biodiversity research involving material gathering, however, scientists typically track samples with *sampleID*, and even different body parts of the same individual can have distinct *sampleID*s. This granularity is lost when data is forced into a single Occurrence table. DwC-DP allows the meaning of identifiers to be preserved by maintaining a clear distinction between Occurrence and Material.

Completeness limitations. The dataset could only be mapped to a limited extent due to completeness issues. Much information was not captured from the start because it was not anticipated for publication, while other details were omitted because DwC-A did not support the information in a structured way. The dataset originally existed as a single Occurrence table and therefore lacked substantial contextual information, which could only be enriched by contacting the data provider. With DwC-DP, such information can now be reported, encouraging researchers to

record these details from the start so they become FAIR rather than remaining buried in manuscripts and/or notes.

Challenges of mapping into DwC-DP. Mapping the dataset into DwC-DP required substantial knowledge of the model, which currently consists of 70 tables and presents a steep learning curve. The presence of many-to-many relationships made it difficult to keep track of connections across tables, especially given the high degree of normalization. Numerous identifiers had to be invented during the process, including many (e.g., *identificationID*) that data providers may have never used. If persistence of such identifiers is expected, updating datasets will remain a challenge, as these IDs are difficult to manage and maintain over time.

Ultimately, publication of the Broke West Fish test dataset as a DwC Data Package required 14 table schemas (**Table 1**).

Table 1. DwC-DP table schemas implemented to comprehensively map all data from the *Broke West Fish test* dataset.

Agent	Material	Protocol
AgentIdentifier	MaterialAssertion	Survey
Event	MaterialMedia	SurveyProtocol
EventAssertion	Media	SurveyTarget
Identification	Occurrence	

Insektmobilen

The Insektmobilen dataset represents a biotic survey with target organisms (insects collected with roof-top-mounted nets, by citizen scientists) and land cover scopes, collection material, media, DNA metabarcoding data, and multiple identifications. The dataset was originally published as an occurrence core Darwin Core Archive. The DwC-DP example dataset folder includes a ReadMe, the originally mapped data used for testing against the database schema, and the output data.

The Insektmobilen data was used to explore:

Information constrained by the DwC-A star schema. Ideally, the original DwC-A dataset would have been shared as a DwC Event core dataset; but, the star schema limitation meant that, if the original data were shared as an Event core dataset, DNA metabarcoding data associated with occurrence data could not have been shared using the DNA-derived data extension. In order to ensure that published occurrence data included relevant DNA metabarcoding data, the dataset was shared as a DwC Occurrence core dataset (**Figure 4**). Sharing the data using the occurrence core meant that the connection between the data was either misrepresented or excessively repeated. For example:

1. Images taken from the material output of the sampling Events are associated with the individual occurrences although they represent bulk insect samples (Material) from which the occurrences are derived.

2. Measurements of the bulk insect sample (i.e. dry weight biomass and DNA concentration) are repeated for each occurrence, although they represent measurements of the subsampled material.
3. DNA sequences, their associated identifications, and sampling Event information were repeated for each occurrence record, although they pertained to either a child or a parent Event.
4. Survey and laboratory analyses yielded several Material entities that are not well-represented in the Darwin Core Archive.

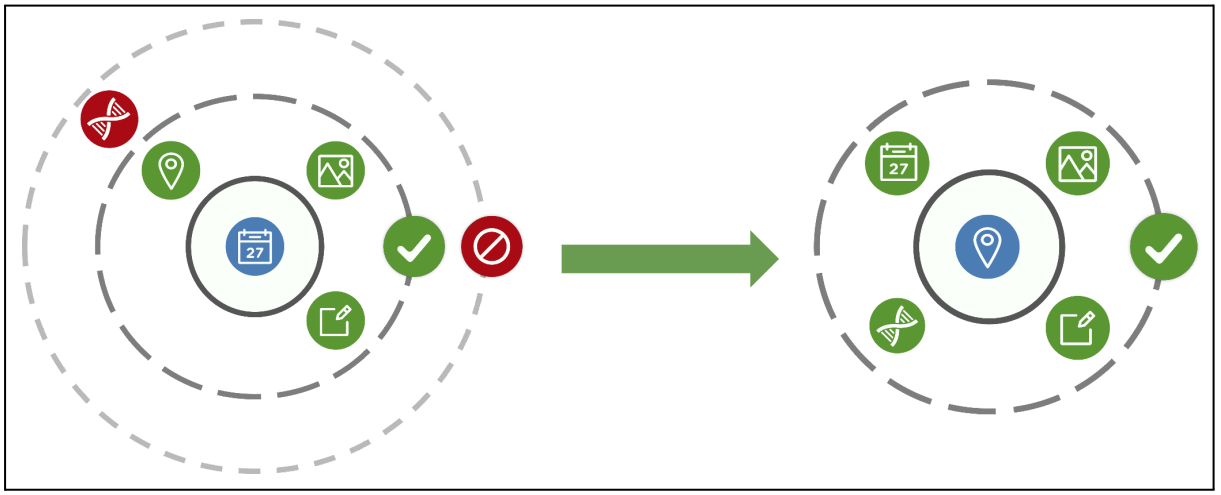

Figure 4. Darwin Core Archive star schema limitation for sharing DNA-derived extension associated with the Occurrences. Only one layer of extensions is allowed and must relate to the core table.
To the left: representation of breaking the star schema if the data were shared with DwC Event core.
To the right: sharing the data with DwC Occurrence core. The date icon represents DwC Event core, the note icon represents the extended measurement or fact extension, the image icon represents the audiovisual media extension, the place marker icon represents occurrences, and the double helix DNA icon represents the DNA-derived data extension.

Mapping DNA metabarcoding survey data to DwC-DP. Mapping the Insektmobilen dataset to DwC-DP provided an opportunity to share a more detailed and accurate representation of data from the project. The process required the implementation of 12 table schemas. A list of these schemas is shared in **Table 2,** and the relationships between these schemas are illustrated in **Figure 5**.

Table 2. DwC-DP table schemas implemented to comprehensively map all data from the *Insektmobilen* dataset.

Event	MaterialAssertion	Occurrence
EventMedia	MolecularProtocol	Survey
Identification	NucleotideAnalysis	SurveyTarget
Material	NucleotideSequence	

In general, information was not repeated excessively and more information could be shared in a structured way, with clear data relationships across tables. As with the newly ratified Humboldt Ecological Inventory extension for DwC Event core datasets, it was possible to define the scope and

targets of the data in the Survey and SurveyTarget tables. Survey-specific information was mostly described in the metadata of the DwC-A version.

This use case also served as a mapping exercise to structure the tables for sharing DNA-derived data: NucleotideAnalysis, NucleotideSequence, and MolecularProtocol. The NucleotideAnalysis table defines which protocol was used to yield a specific DNA sequence, and the sequence's relative abundance, from a specific material of a sampling Event. It enabled linking the sequence to the purified DNA extract in a collection, while the Material table provided information on the bulk sample, the raw DNA extract, and the purified DNA extract, thus increasing potential reuse of not only the data point, but the breadth of material behind it. Each DNA sequence was shared only once in the NucleotideSequence table, and the Identification table allowed for sharing the identification of each DNA sequence, but also any morphological identifications of the same material, from the same event. Mapping the data to DwC-DP solved several problems faced when mapping the data as a DwC-A:

1. **Sharing media of an entire research project**. Media can be associated with all the main tables. Currently, media of the material is shared, but videos and images taken by the citizen scientist can potentially be shared as EventMedia as well.
2. **Limit repetition of DNA-derived identifications**. Only sharing DNA sequences once allows for much smaller tables compared to sharing them as individual Occurrences. This is particularly useful for eDNA surveys that yield thousands to millions of sequences.
3. **Highlight data provenance**. Separating materials from Events allows for a clear distinction between survey Events and subsequent material handling in a structured manner.

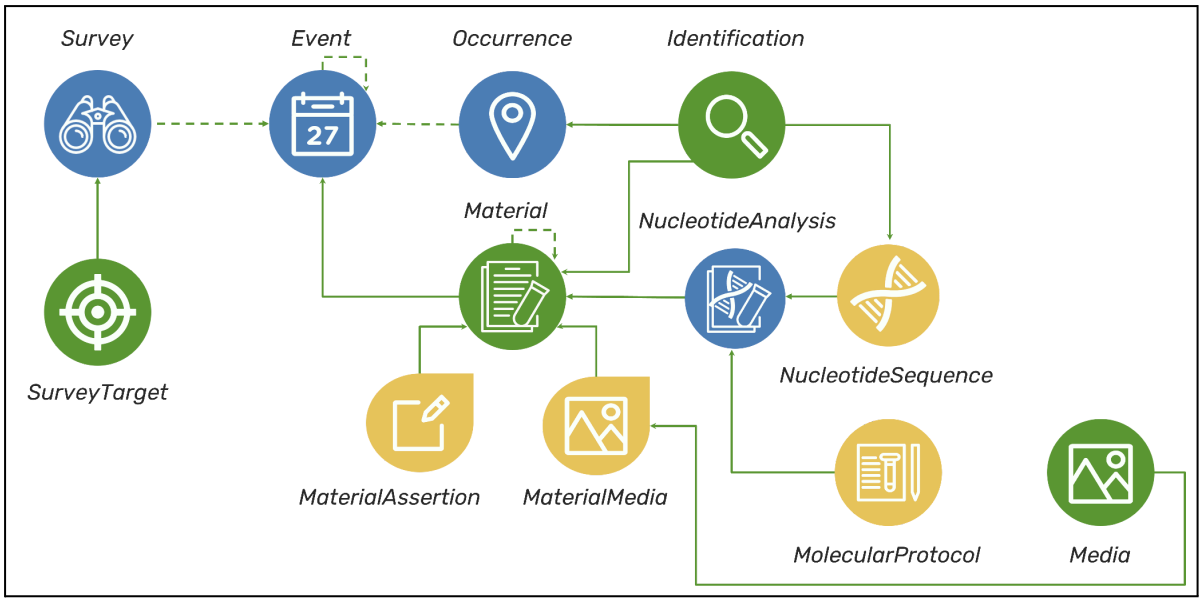

Figure 5. Tables included in the DwC-DP mapping of the Insektmobilen dataset and how they are connected.

Completeness limitations of the project data. DNA sequences had not been annotated for two out of three primer pairs, which did not allow for proper testing of identifications coming from multiple protocols and primers. The output tables of the use case have some pending updates to be made. For example, the Occurrence table contains missing occurrence IDs, and the eventID is missing for the morphological identifications. Because of the missing occurrenceIDs, the dataset is not published

on the test IPT yet. Sharing richer information on Agents was purposefully left out of the mapping due to privacy concerns. However, several Agent-related tables could be included to highlight when different agents were involved in data generation or processing, and acknowledge the contribution of the citizen scientist who participated in the project.

Challenges of mapping into DwC-DP. Although identifications were only shared for one primer pair, the same species could be detected with multiple primers, which potentially could lead to duplicated occurrences. If the dataset is updated with identifications from the two other primers used, it would require the publisher to choose a *isAcceptedIdentification* e.g., for the reference database match with the highest confidence or another threshold. Although much more information could be shared in a structured way in DwC-DP compared to DwC-A, creating identifiers for all tables was challenging. Most of the identifiers that had to be created were constructed from a data context and would probably have benefited from being constructed as non-meaningful identifiers. However, constructing the identifiers based on the data helped inform the relationship of the IDs at a time when the relationship explorer and the quick reference guide were not available. Lastly, structuring records from a highly nested sample pool into the available tables also proved to be a challenge, although an increased familiarity with the model as mapping progressed helped provide structure eventually. The main challenge was to figure out what constituted an event; however, this is also a challenge when constructing a DwC-A, hence not a unique issue for DwC-DP. The Material table helped split any subsampling into separate categories, and the option of referring to whether the material was *derivedFrom* or *isPartOf* allowed for clear data provenance tracking.

Other community feedback – GitHub

Feedback from the biological data publishing community was also obtained via two GitHub issue trackers: the general DwC-DP GitHub issue tracker and the DwC-DP early community testing issue tracker. These two avenues for reporting issues and discussing concepts, terms, and applications successfully drove a broad series of debates. Feedback and discussion in GitHub included topics such as:

Versioning and addition of new schemas. The community should be able to extend the data model through the addition of new schemas. DwC Task Groups could be a mechanism to develop and test each new schema proposed. Each new schema would need to respect the DwC-DP profile (see discussion in issue #114) to ensure there are no adverse effects on existing table schemas. The DwC Maintenance Group will provide support to integrate new schemas, resulting in new DwC-DP versions.

Seeming duplication of classes. It was noted that the DwC-A resource relationship concept was retained even though DwC-DP includes an OrganismInteraction class. This is because the resource relationship concept can support linkages that cannot be otherwise supported by the OrganismInteraction class.

Reuse of DwC Properties and Categories. Inquiries were made into the fact that sometimes DwC terms are reused and adopted into DwC-DP, but not always. While the aim in developing DwC-DP was to reuse Darwin Core terms as much as possible, there is a balance in determining where to

retain existing class or term names and when to create new names. For details, see the GitHub issue #113 - Table Schema property proposal - terms in context. For example, terms in DwC Extended Measurement or Facts extension (EMoF) are integrated into DwC-DP as Assertions. However, DwC-DP adds additional terms in Assertions to validate or add contextual meaning, for example, referencing controlled vocabularies used for data mapping (issue #89).

eDNA and metabarcoding data. Concerns were raised about creating new schemas and terms for DNA data where other standards exist. The DNA-relevant tables in DwC-DP were constructed based on terms included in the DNA-derived data extension maintained by GBIF, which itself is constructed by incorporating existing data standards, mainly MiXS. An output format for DwC-DP has been implemented in the GBIF Metabarcoding Data Toolkit to help with testing.

GeologicalContext cardinality. The reviewer identified places where the schema table connections required re-evaluation. The issue was discussed thoroughly in GitHub and addressed within the model. See DwC-DP issue #108 for more details.

Taxon class. A Taxon Class is currently not included in DwC-DP public review as it requires alignment with ongoing community work. See DwC-DP issue #88 for more details.

Phylogenetic trees and phylogenetic tree tips. To support phylogenies without supporting Taxon explicitly is hard to justify. With that reasoning, PhylogeneticTree and PhylogeneticTreeTip will not be part of the DwC-DP going to the original public review (see issue #111).

Unresolved issues/future challenges

The DwC Maintenance Group, or task groups formed under the interest group, will work on potential future integration of references, taxonomy, phylogeny, provenance, and rights.

Taxon class. The first iteration of DwC-DP to be presented for ratification does not include a Taxon class. Potential integration of Taxon-level information into DwC-DP awaits developments to integrate and standardize the Catalogue of Life Data Package (CoL-DP) and the Taxonomic Concept Schema version 2 (TCS2). See DwC-DP issue #88 for more details.

Phylogenetic trees. Early plans aimed to support the sharing of phylogenetic trees and phylogenetic tree tips with DwC Data Packages. Efforts to enable this feature will remain on hold until other taxon-related issues have been addressed. Refer to DwC-DP GitHub issue #111 and the original discussion in issue #10.

Rights and Provenance. As a result of discussion on GitHub and in other forums, both UsagePolicy (GitHub issue #119) and Provenance (GitHub issue #103) will be included as classes in DwC-DP.

Conclusions

After several years of discussion, debate, and stress testing with myriad data types and formats, we feel confident in pushing the Darwin Core Data Package forward for community review. DwC-DP grew from individual use cases, integrating each one while ensuring the whole system worked for all. It

addresses the need to capture much more diverse and complex biological data types, but responds to community demands to limit the number of publishing models and to ensure as much consistency with Darwin Core Archives as possible. DwC-DP is compatible with Darwin Core and supports all non-checklist use cases without the limitations of the star schema.

Acknowledgements

We are grateful for the many individuals, groups, and organizations that have contributed to the development of the Darwin Core Data Package.