Use Case: Global Malaise Programme

Status: In review

Summary

As part of the Global Malaise Programme, organisms, their remains, and media are collected and identified with a set frequency over a year-long period in each of many locations. Environmental conditions may also be captured with these events and potentially at a higher frequency.

- The Global Malaise Programme consists of an array of one trap at each of 153 sites. The entire Programme can be captured as an encompassing **Event**, using the geographic and temporal bounds of all individual **Events** related to it.
- The **Location** for a single trap was (often) selected for convenience rather than for any systematic biological criteria.
- Collecting **Events** normally happen once a week, extracting the accumulated biomass from the alcohol receptacle for the period since the previous collection as a **MaterialSample**.
    - Measurements of temperature, humidity, soil moisture and trap-adjacent rainfall and wind speed/direction are gathered every five minutes on the site.
    - A **StillImage** of the entire MaterialSample spread on a sorting tray is captured.
    - A **StillImage** cropped from the full-tray **StillImage** may be created for an **Organism**.
    - **A StillImage** taken under a microscope may be created for an **Organism**.
    - Sets of **StillImages** of **Organisms** will be captured in iNaturalist Occurrence records.
- **MaterialSamples** are sent to the Centre for Biodiversity Genomics (CBG) in Guelph where they are weighed.
    - A subset of **Organisms** is extracted from each **MaterialSample,** curated into the CBG **Collection**, and put in 96-well plates where **StillImages** are taken of them.
    - Plates go through the CBG sequencing pipeline, in which **MaterialSamples** of DNA are extracted from the **Organisms** and stored in the CBG Collection.
    - DNA **Sequences** from the COI region are extracted from **MaterialSamples** from the CBG **Collection**.

    - Collecting **Event** and **Specimen** metadata (**Location**, date, collector), a rough **Identification** (typically family/subfamily/tribe, but sometimes order), the **StillImage** in the well, the Sequences, and trace files are published to the Barcode of Life Data (BOLD) System.
- A future alternative pathway to the previous point is to process the **MaterialSample** for metabarcoding. This would be similar to the BioWide Use Case.
- **Identifications** for **Occurrence** records in iNaturalist and in BOLD may be updated based on new evidence (including updating BOLD from iNaturalist identifications and vice versa).
- iNaturalist record identifiers and links to **StillImages** may be added to BOLD records, if possible.
- BOLD record identifiers may be added to iNaturalist records, if possible.

Highlights

MaterialSamples captured as StillImage-based Observations with Identifications in iNaturalist and as parallel (duplicate Occurrence) MaterialSamples with StillImages, DNA subsamples, DNA sequences, and Identifications to Operational Taxonomic Units in BOLD. Similar to the BioWide Use Case in that Identifications should be updatable from sources (sequence reference libraries) external to the primary dataset. Presents a challenge of tracking individual organisms between source and BOLD as well as Identifications between BOLD and iNaturalist. This case has the potential to integrate with phylogenies.

Concepts - see Glossary

Conceptual Model

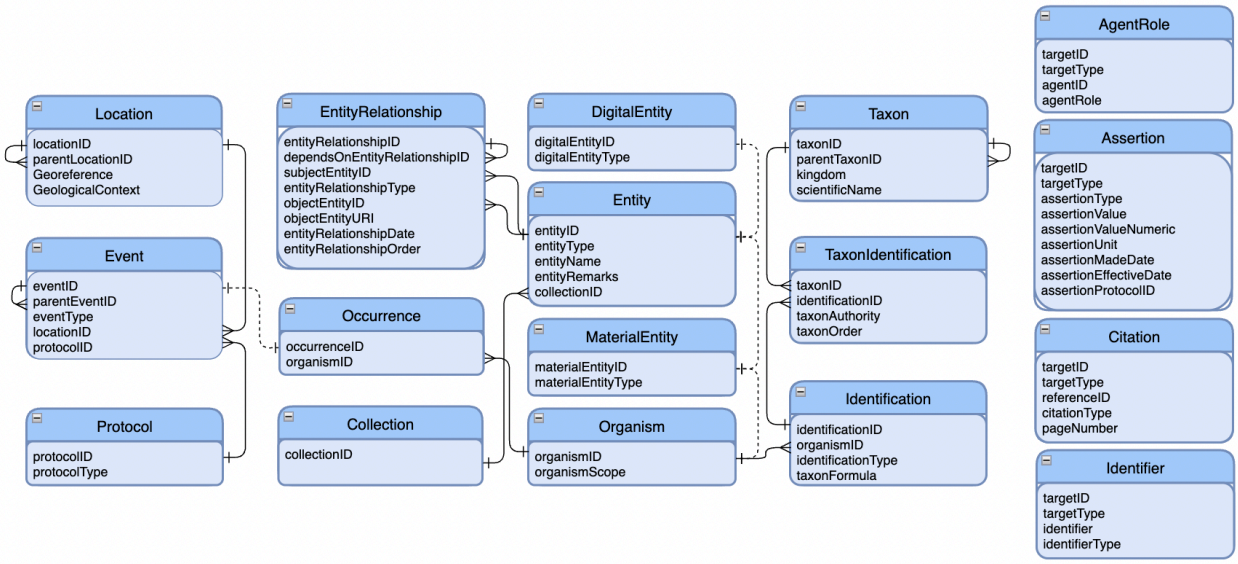

Figure 1: A conceptual model (2024-04-04) covering the activities outlined in the Global Malaise Programme use case, which includes periodic sampling in selected locations, imaging of individual **Organisms** for publication in iNaturalist, imaging and processing of some of the same **Organisms** for DNA extraction and **GeneticSequences** in BOLD, and **Identifications** from both iNaturalist and BOLD with the desire to link the **Organism** records and **Identifications** between those two **Collections**.

The project can be represented as a parent-most **Event**, the spatial and temporal limits of which encompass all of the **Events** within it. The project can consist of a set of collecting **Events** that happen at one or more **Locations** during one or more periods of time (included in the **Event**), for example, the 52 weekly collecting **Events** can have the one-year project at a **Location** as their parent **Event**. Each collecting **Event** may follow a specified **Protocol** with respect to the target material sample (a **MaterialEntity**). **Organisms** (also **MaterialEntities**) may be isolated

from the material sample (for example, by taking photos, each of which is a **DigitalEntity** that represents a **MaterialEntity** - an **Organism** in this case) as its subject. Each **MaterialEntity** and **DigitalEntity** can optionally be added to a **Collection**.

A genetic sequence (a **DigitalEntity)** is *derivedFrom* a DNA extract (a **MaterialEntity**), which in turn is *derivedFrom* an **Organism** or part thereof (a **MaterialEntity**) that was *derivedFrom* a material sample (a **MaterialEntity** consisting of a mixed set of arthropods in alcohol), which was collected during the **Event** at the **Location** and during the period when a collection **Protocol** was executed.

Events can be created for specific or summarized environmental data at whatever temporal granularity is desired - they do not have to be only collecting **Events**. Environmental data would be captured in Assertions (described in Common Models).

Photos (**DigitalEntities**) in iNaturalist can be used as evidence for multiple **TaxonIdentifications** of an **Organism**.

GeneticSequences (each a type of **DigitalEntity**) in BOLD can be used as evidence for multiple **TaxonIdentifications** of an **Organism**. A **SequenceTaxon** establishes an **Identification** from a genetic sequence as having an **Organism** of a particular **Taxon** (BIN, and OTU) as a source. The **Taxon** identified will depend on the sequence reference library, which can change over time, and which may or may not be an accepted **Taxon** in the GBIF Backbone Taxonomy.

An arbitrary number of **Assertions** can be made about each class. **Assertions** can be quantitative or qualitative and can have **Assertions** made about them as well. **Agents** can have roles with respect to any class as well, including **Assertions**. Instances of any class may be referenced in **Citations** and have alternate **Identifiers**. These four common aspects of all use cases can be found in Common Models.

Publishing Model - iNaturalist

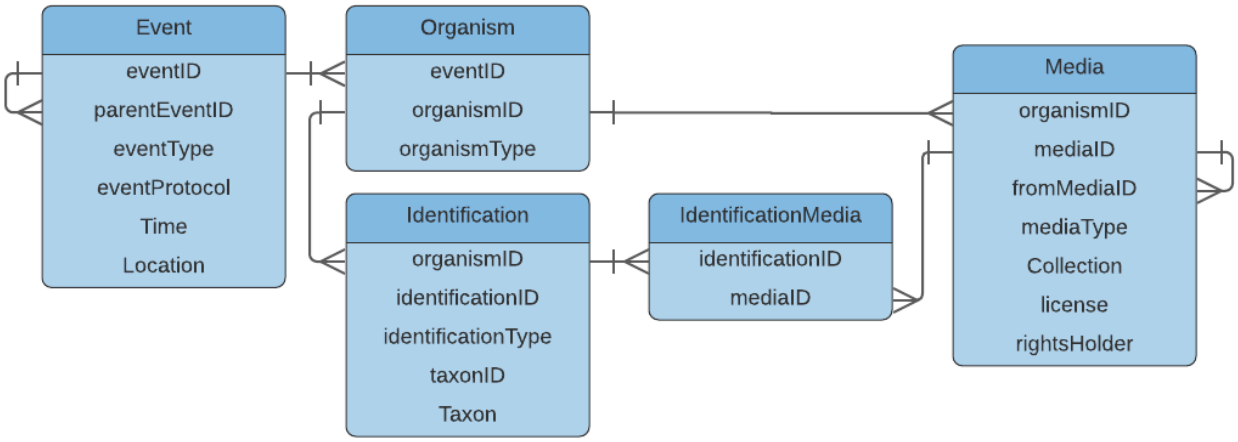

Figure 2. A simplification of the part of the conceptual model for the purpose of publishing Observation data from iNaturalist. For an explanation of this model, see Use Case: iNaturalist Observations.

Publishing Model - BOLD

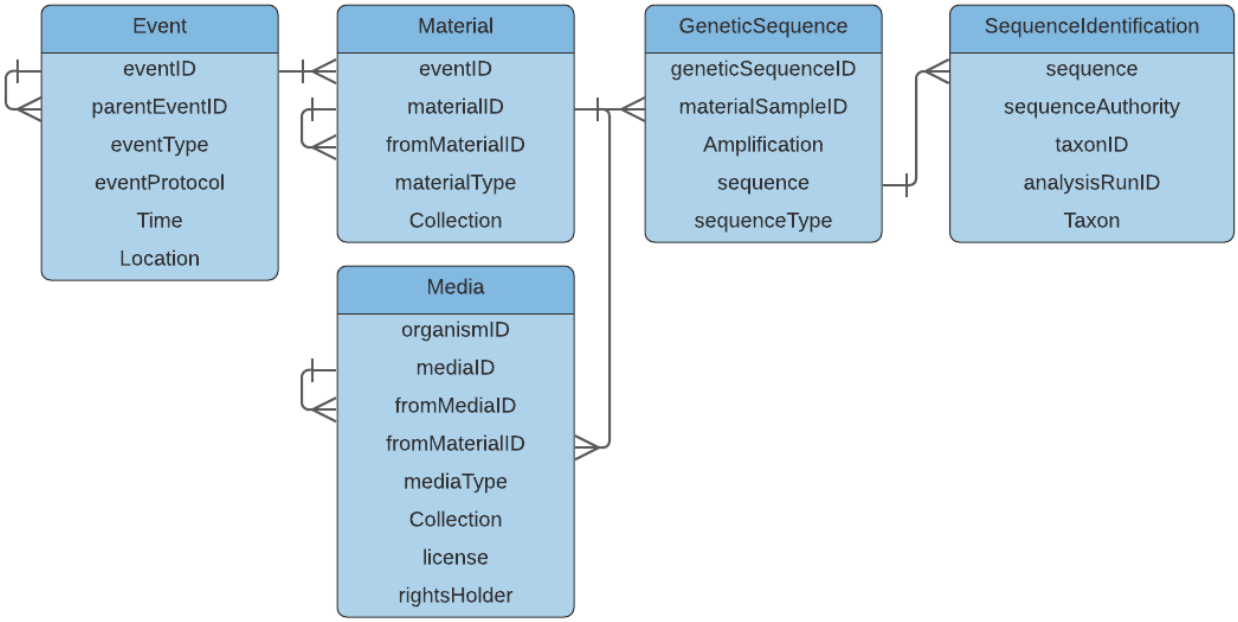

Figure 3. A simplification of the part of the conceptual model for the purpose of publishing genetic sequence-based identifications from BOLD.

Event - one row for each **Event** in which an **Organism** or part of an **Organism** was collected, one row for the DNA extraction from an **Organism** or part of an **Organism**, and optionally one row for the creation of each instance of digital **Media** having an **Organism** or derivative **Material** as a subject. The Darwin Core Event Core could be used with the broadening of the

dwc:samplingProtocol attribute to be any kind of *eventProtocol*, such as a *mediaCreationMethod* and *extractionMethod*.

Terms - *eventType, eventProtocol* and potentially all terms in the most recent version of Event Core (https://github.com/gbif/rs.gbif.org/tree/master/core).

Material - one row for each **Organism** or part of an **Organism**, including DNA extractions. This table could have **Collection** information in it, or a **Collection** table could be used separately and linked.

Terms: *materialID* (identifier for the **Organism** or part of the **Organism**, including the DNA extraction), *fromMaterialID*, *materialType (~dwc:preparations)*, *preservationMethod*, *hasDNA*, *availability* (=*dwc:disposition*), *condition*, *catalogNumber*, *recordNumber*, *recordedBy*, *recordedByID*, *otherCatalogNumbers*, *materialSampleRemarks* and **Collection** information, either integral to the extension or external via a reference.

Media - one row for each digital media entity (StillImage) derived from an **Organism**. This extension could have **Collection** information in it, or a **Collection** table could be used separately and linked. The Audiovisual Media Description extension should be used for this case (https://tools.gbif.org/dwca-validator/extension.do?id=http://rs.tdwg.org/ac/terms/Multimedia).

Terms: *materialID* (identifier for the **Organism** that was the subject of the **Media**), *mediaID*, *fromMediaID*, *mediaType (~dc:type)*, *license*, *rightsholder*, and **Collection** information, either integral to the extension or external via a reference.

GeneticData - one row for each DNA extract. Though the extracts are also **MaterialEntities**, the current data publishing schema merges that with amplifications and resulting sequences. I am not trying to unravel that mess here.

Terms: all terms from the latest version of DNA-derived Data Extension (https://tools.gbif.org/dwca-validator/extension.do?id=http://rs.gbif.org/terms/1.0/DNADerivedData).

SequenceIdentification - one row for each **Taxon Identification** from a sequence in a sequence reference catalog.

Terms: *sequenceAuthority*, DNA sequence (from DNA-derived Data Extension), all **Identification** terms, all **Taxon** terms necessary for **Identifications**.

References

Common Models
Use Case: iNaturalist Observations
Use Case: BioWide eDNA
https://biodiversitygenomics.net/projects/gmp/
https://github.com/gbif/rs.gbif.org/tree/master/core
https://rs.gbif.org/extension/
https://tools.gbif.org/dwca-validator/extensions.do
https://tools.gbif.org/dwca-validator/extension.do?id=dwc:Event
https://rs.gbif.org/core/dwc_event_2016_06_21.xml
https://tools.gbif.org/dwca-validator/extension.do?id=dwc:MaterialSample
https://github.com/gbif/rs.gbif.org/blob/master/sandbox/core/dwc_material_sample.xml

https://tools.gbif.org/dwca-validator/extension.do?id=dwc:Identification
https://rs.gbif.org/extension/dwc/identification.xml
https://tools.gbif.org/dwca-validator/extension.do?id=http://rs.tdwg.org/ac/terms/Multimedia
https://github.com/gbif/rs.gbif.org/blob/master/extension/ac/audubon_2020_10_06.xml
https://tools.gbif.org/dwca-validator/extension.do?id=http://rs.gbif.org/terms/1.0/DNADerivedData
https://github.com/gbif/rs.gbif.org/issues/63

Grames EM, Montgomery GA, Boyes DH et al. (2022) A framework and case study to systematically identify long-term insect abundance and diversity datasets. Conservation Science and Practice e12687. https://doi.org/10.1111/csp2.12687

Acknowledgements

Donald Hobern