WorldFAIR

| | |
|---|---|
| Project Title | Global cooperation on FAIR data policy and practice |
| Project Acronym | WorldFAIR |
| Grant Agreement No | 101058393 |
| Instrument | HORIZON-WIDERA-2021-ERA-01 |
| Topic, type of action | HORIZON-WIDERA-2021-ERA-01-41 HORIZON Coordination and Support Actions |
| Start Date of Project | 2022-06-01 |
| Duration of Project | 24 months |
| Project Website | [http://worldfair-project.eu](http://worldfair-project.eu) |

# D9.1 Data standard for sharing ecological and environmental monitoring data documented for community review

| | |
|---|---|
| Work Package | WP9 - Biodiversity |
| Lead Author (Org) | Joe Miller (GBIF) |
| Contributing Author(s) (Org) | Tim Robertson (GBIF), John Wieczorek (GBIF contractee) |
| Due Date | 30.05.2023 |

DRAFT NOT YET APPROVED BY THE EUROPEAN COMMISSION

| | |
|---|---|
| Date | 17.05.2023 |
| Version | 1.0 DRAFT NOT YET APPROVED BY THE EUROPEAN COMMISSION |
| DOI | https://doi.org/10.5281/zenodo.7849241 |

Dissemination Level

| | |
|---|---|
| X | PU: Public |
| | PP: Restricted to other programme participants (including the Commission) |
| | RE: Restricted to a group specified by the consortium (including the Commission) |
| | CO: Confidential, only for members of the consortium (including the Commission) |

## Versioning and contribution history

| Version | Date | Authors | Notes |
|---|---|---|---|
| 0.9 | 21.04.2023 | Joe Miller (GBIF) | Draft for internal review |
| 1 | 17.05.2023 | Joe Miller (GBIF) | Final, post internal review |
| | | | |
| | | | |

## Disclaimer

WorldFAIR has received funding from the European Commission’s WIDERA coordination and support programme under the Grant Agreement no. 101058393. The content of this document does not represent the opinion of the European Commission, and the European Commission is not responsible for any use that might be made of such content.



‘Global cooperation on FAIR data policy and practice’ (WorldFAIR) has received funding from the European Union’s Horizon Europe project call HORIZON-WIDERA-2021-ERA-01-01, grant agreement 101058393.

## Abbreviations and Acronyms

| | |
|---|---|
| CBD | Convention on Biological Diversity |
| DDI | Data Documentation Initiative |
| DwC | Darwin Core |
| FAIR | Findable, Accessible, Interoperable, Reusable |
| FIP | FAIR Implementation Profile |
| GBIF | Global Biodiversity Information Facility |
| TDWG | Biodiversity Information Standards |



'Global cooperation on FAIR data policy and practice' (WorldFAIR) has received funding from the European Union's Horizon Europe project call HORIZON-WIDERA-2021-ERA-01-01, grant agreement 101058393.

## Executive Summary

Biodiversity standards are essential for FAIR data, in particular for interoperability. Current standards need to be improved with new data models to better reflect the complexity of biodiversity and serve the information needed to address biodiversity loss and climate change.

This Deliverable D9.1, focused on Task 9.1, describes the FAIR data model being developed in WorldFAIR WP09 with the Global Biodiversity Information Facility (GBIF) leading a community collaboration.

Facilitated by the WorldFAIR project, GBIF's engagement with the biodiversity community has led to this Deliverable – a new draft core Unified Model. The model was developed in collaboration with the Biodiversity Information Standards Group (TDWG) and through community consultation via webinars, open drafting of documents and solicitation of test datasets from the various stakeholders. A review of comparable standards has led to the development of a draft core model framework that is known to the community which should make adoption easier.

The new model is centred around the 'Event' – something happened at some place during some period of time, optimally described by a protocol. This conclusion is based on research which describes how successful models are expressed and the flexibility of the Event to accommodate many types of data. The Unified Model is applicable to all currently-used data types and potential new data to be shared with GBIF. The current community engagement approach (more on which in D9.2) is to test individual components of the model with engagement activities and example datasets. This will continue with new tests expanding the potential utility of the model.

The WP tasks performed to date (collation of previous material, engagement with TDWG and Darwin Core (DwC) standard leads, webinars, use of shared documents to build use cases, building of exemplar datasets for collection management systems, and provision of several avenues for feedback) have resulted in the new provisional model currently under consultation. Testing and community engagement to date indicates that the model will better reflect the complexity of biodiversity data leading to more efficient use of our community's data in research and policy. The feedback we have received also indicates a steep learning curve for the future implementation of the data model. This feedback is essential for the development of publishing tools.

This work aligns with the overall objectives of WorldFAIR by focused development on improving the **interoperability** of biodiversity data. Our FAIR Implementation Profile (FIP) will be enhanced by this improved functionality. This work promotes cross-domain interaction, as the Unified Model will enhance sharing of data in related Work Packages such as Agricultural Biodiversity, Oceans, and Geochemistry in the final portion of the WorldFAIR grant period. This work has been undertaken in alignment with the overall WorldFAIR goals, in particular WP02 on Engagement, Synthesis, Recommendations and FAIR Assessment.



'Global cooperation on FAIR data policy and practice' (WorldFAIR) has received funding from the European Union's Horizon Europe project call HORIZON-WIDERA-2021-ERA-01-01, grant agreement 101058393.

# Table of contents





'Global cooperation on FAIR data policy and practice' (WorldFAIR) has received funding from the European Union's Horizon Europe project call HORIZON-WIDERA-2021-ERA-01-01, grant agreement 101058393.

## 1. Introduction: The need for a more expansive model

Biodiversity knowledge is critical to sustain our planet and this was emphasised in the recent COP15 Kunming-Montreal Global Biodiversity Framework (GBF). National governments need data to understand biodiversity and to report on their obligations to the GBF. This requires many types of data including data on the distribution of biodiversity, most often assessed by the distribution of species.

The Global Biodiversity Information Facility (GBIF) is called out in the GBF Target 21 as an complimentary indicator to provide a measure of growth in accessible species occurrence records. The evidence of a species occurrence can vary from a human observation, a museum collection, or machine observations. Therefore, as a global infrastructure GBIF must make evidence from various sources interoperable, perhaps the most challenging aspect of FAIR data. The biodiversity crisis accentuates the need for more and more detailed data — a challenge to GBIF.

Darwin Core (DwC) — the most commonly used data standard in the GBIF community—has provided a simple and effective framework for supporting the growth of species occurrence data available from the GBIF network (Wieczorek et al. 2012). DwC (see Figure 1) allows easy integration of these varied data types and its use has been critical for the GBIF community to build evidence of 2.3 billion species occurrence events since 2001. Its simplicity is what permits interoperability of data from over 83,000 datasets from 2,000 publishers from nearly every country. The capacity and incentive to share data varies and DwC has allowed GBIF to grow rapidly to be useful for research and policy (Heberling et al. 2021). It is largely due to the simplicity of publishing using Darwin Core Archives in conjunction with the GBIF Integrated Publishing Toolkit.



'Global cooperation on FAIR data policy and practice' (WorldFAIR) has received funding from the European Union's Horizon Europe project call HORIZON-WIDERA-2021-ERA-01-01, grant agreement 101058393.

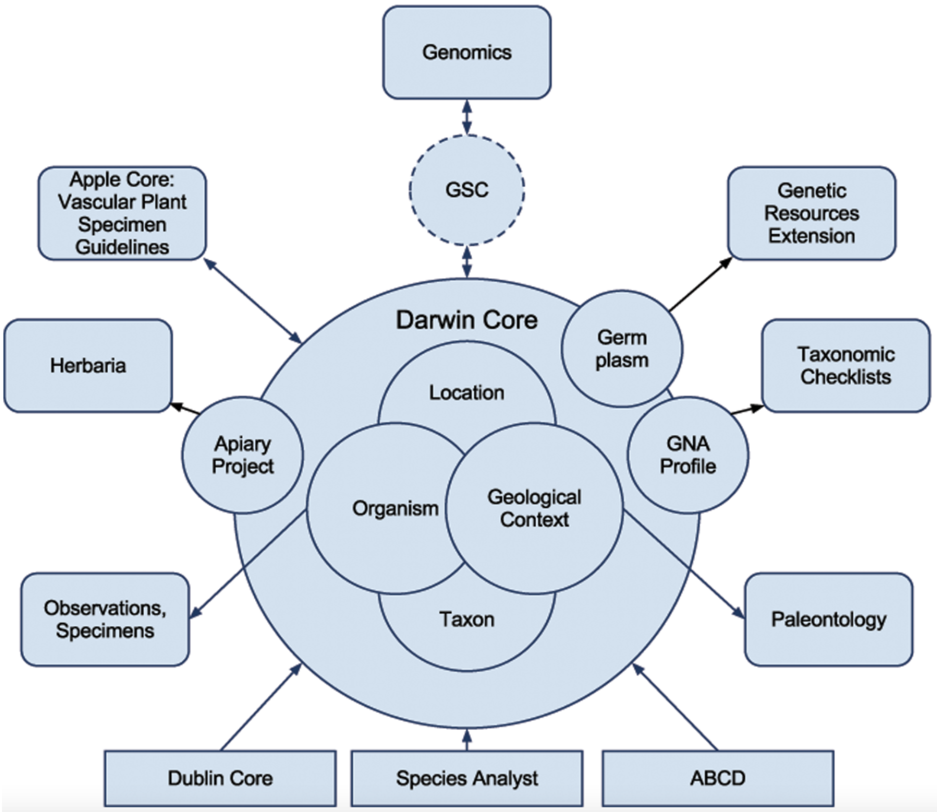


*Figure 1. Darwin Core Start Schema (Wieczorek et al. 2012).*

But the simplicity of the DwC standard, established and maintained by Biodiversity Information Standards (TDWG), and in particular the common Darwin Core Archives data packaging format, has significant limitations when it comes to shaping data from diverse sources. These limitations are centred around the 'flat' DwC core set of records to which extensions can be connected only one level deep. Firstly, DwC results in an oversimplification which sometimes results in the loss of data richness, which in turn can affect compliance with the FAIR principles. Secondly, DwC may transform data incorrectly based on its limited focus on the taxon and the occurrence.

The Darwin Core community clearly understands this tension between ease of use of publication, against the loss of information in an oversimplified model and has encouraged expansion of data publishing capabilities to supply missing semantic rigour. The GBIF 20-year review (CODATA 2020) recommended maintaining and building on its status as the most comprehensive source of biodiversity data. The CODATA review acknowledges the issues associated with the complexity of biodiversity data and identifies the need to accommodate more varied data types to meet growing



'Global cooperation on FAIR data policy and practice' (WorldFAIR) has received funding from the European Union's Horizon Europe project call HORIZON-WIDERA-2021-ERA-01-01, grant agreement 101058393.

research needs.

## 2. A case-study driven approach

In the new *Strategic Framework 2023-2027* GBIF states that "biodiversity data is more complicated than 'just' the occurrence of species in time and space: organisms interact, co-occur, move and evolve (GBIF 2021). This implies a need for richer and more varied types of data than GBIF has thus far supported" (p. 2). This sets the course for a long-term initiative focused around expanding the DwC data model.

In order to meet the challenge to improve our data model, GBIF has developed an approach to working closely with the publishing and the Darwin Core communities in an interactive manner. The work is more than can be accomplished in the WorldFAIR timeline. The work is expected to take several years and will be enhanced by community participation and knowledge gained through trial and error.

GBIF's approach is to solicit narratives of use cases that are known to represent novel data publishing challenges today. These narratives are developed in use case templates which are then transformed into an interpreted summary (see Figure 2). From the summary narrative the key components of the concept model are identified, such as Location, Event, Material, Media, and Taxon. Critically, the relationships among the concepts are an essential component to understand. Finally, the conceptual model is simplified to the necessary components required in a publishing model. To date, twenty-three narrative case studies have been initiated. They vary in depth of development but all were initiated because they bring additional new elements to the system.

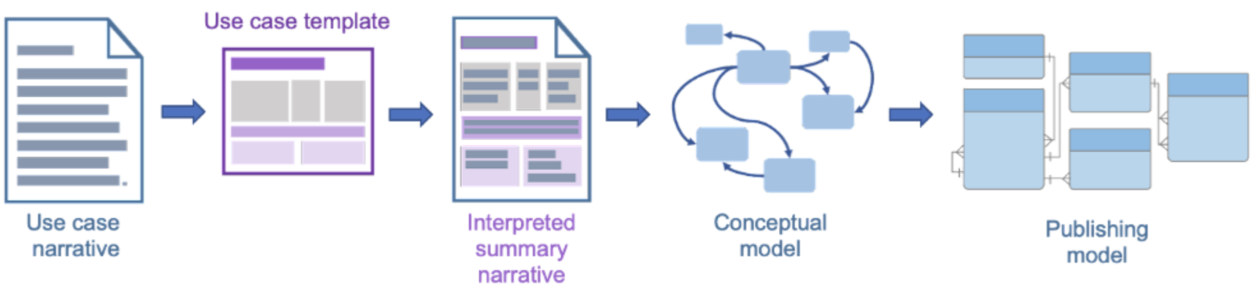


*Figure 2. Diagram showing the progression of the Unified Model development process.*

Each case study includes the construction of a conceptual model, presented as a Unified Modelling Language (UML) Entity-Relationship (ER) diagram. There are several modelling frameworks that



'Global cooperation on FAIR data policy and practice' (WorldFAIR) has received funding from the European Union's Horizon Europe project call HORIZON-WIDERA-2021-ERA-01-01, grant agreement 101058393.

could be used to illustrate the conceptual model, but ER was chosen for two reasons. Firstly, it is a representation that is most likely to be readily understood by the broadest audience, including scientists, developers and data administrators, and secondly the design closely resembles the structure of the data store under which data are to be managed.

## 3. The Event: Initial observations on the Unified Model

Currently DwC is framed around the Occurrence – the occurrence of a species at a particular place and time. The new model is centred around the Event – something happened at some place during some period of time, optimally described by a protocol.  This conclusion is based on preliminary research on how successful models are expressed and the flexibility of the Event.

Scientific physical entities, observations, processes and the relationships between them can be successfully expressed in a wide range of high-level foundational models and ontologies; however, research into this data model literature pointed to the concept of the event as a central organising structure. This includes: the philosophy of events (Casati and Varzi 2020), the Suggested Upper Merged Ontology (SUMO; Niles and Pease 2001), the Sowa Knowledge Representation Ontology (Sowa 2000), Observations and Measurements (OGC O&M; Cox 2011), the Semantic Sensor Network Ontology (SSN and SOSA; Haller et al. 2017), the CIDOC Conceptual Reference Model (International Organization for Standardization [ISO] 2014) and the PROV Data Model (PROV-DM; Lebo et al. 2013). Additionally, process-based Events are foundational features to the current design of the Biological Collections Ontologies (BCO; Walls et al. 2014). Finally the use of Event in the Unified Model also makes it completely compatible with Darwin SW (Baskauf & Webb 2016, Baskauf & Sachs 2018).

The event-based conceptual model (see Figure 3) is extremely flexible, as it allows for capturing data that ultimately helps understand why a specimen was observed in a certain place at a certain time. It is currently used in conjunction with DwC to publish rich hierarchical, ecological datasets, such as the Dutch Vegetation Database. Most importantly it has shown to be effective while used as the foundation for every use case encountered to date.



'Global cooperation on FAIR data policy and practice' (WorldFAIR) has received funding from the European Union's Horizon Europe project call HORIZON-WIDERA-2021-ERA-01-01, grant agreement 101058393.

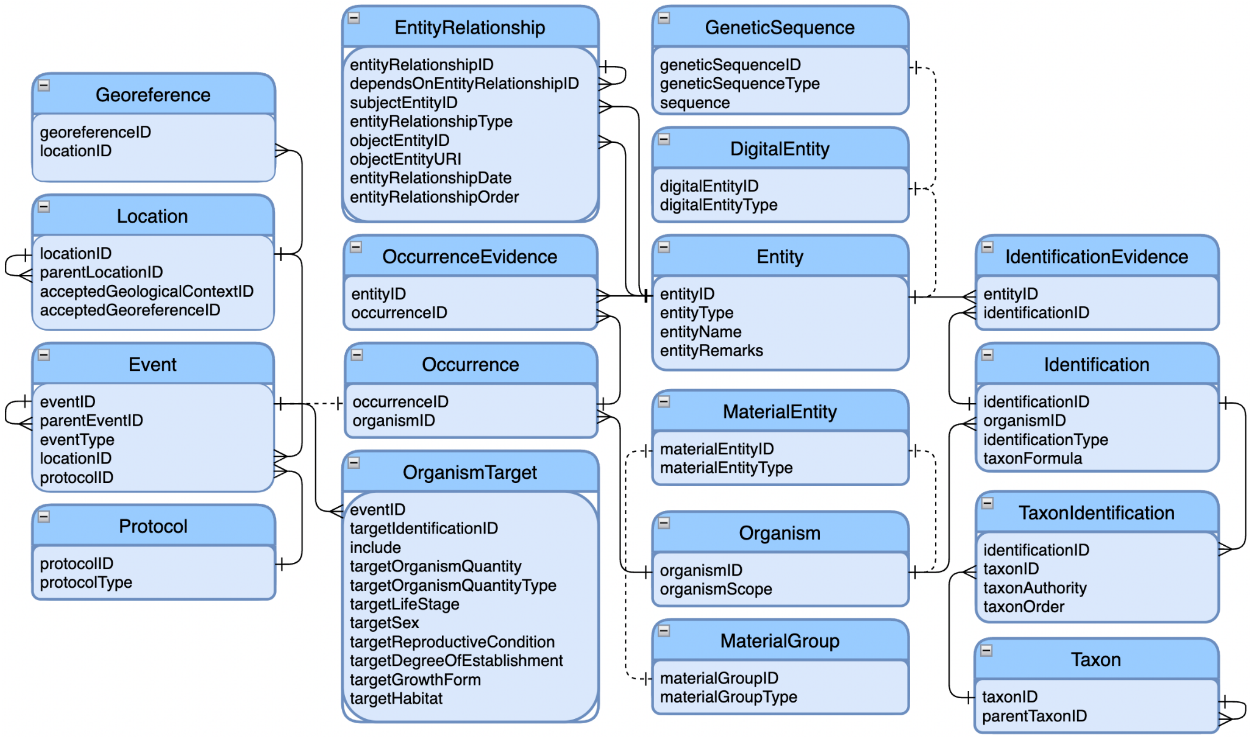


*Figure 3. ER Diagram of the basic concepts of the Unified Model, version 2023-02-09.*

Two other important realisations from the work thus far will be of particular interest to data publishers. Firstly, the complexity of the conceptual model can be distilled into a less complex data publishing model and in many cases existing publishing schemas will be sufficient. Secondly, it is clear that user support, carefully documented and tailored to a family of similar use cases, will greatly benefit the uptake of new data publishing models. Therefore, GBIF has worked to build successful prototype implementations covering a subset of carefully reviewed use cases. This practical experience will aid the Darwin Core Maintenance Group to foster changes in the Darwin Core standard to put this work into practice.

The details of the Unified Model are built around the following four foundational classes (see Figure 4), whose origins come from the PROV-O ontology and Darwin Core standard:

- An **Entity** is a physical, digital, conceptual, or other kind of thing with some fixed aspects; entities may be real or imaginary. An organism, a digital photo and a taxon are examples of Entities that are physical, digital and conceptual respectively.



'Global cooperation on FAIR data policy and practice' (WorldFAIR) has received funding from the European Union's Horizon Europe project call HORIZON-WIDERA-2021-ERA-01-01, grant agreement 101058393.

- An **Event** is a process, action or activity that occurs at some location during some time and acts upon or with entities. A specimen being collected, a species determination and a DNA extraction are simple examples of Event.
- An **Agent** is a person or thing that takes an active role in an Event or bears responsibility for an entity existing. A person, an organisation or an automated sensor are examples of Agent. Note that an Entity may also take the role of an Agent in some scenarios.
- An **Assertion** is a record of an annotation, observation, or measurement made by an Agent at some time, optionally following a described protocol. These are essentially measurement or observation Events.

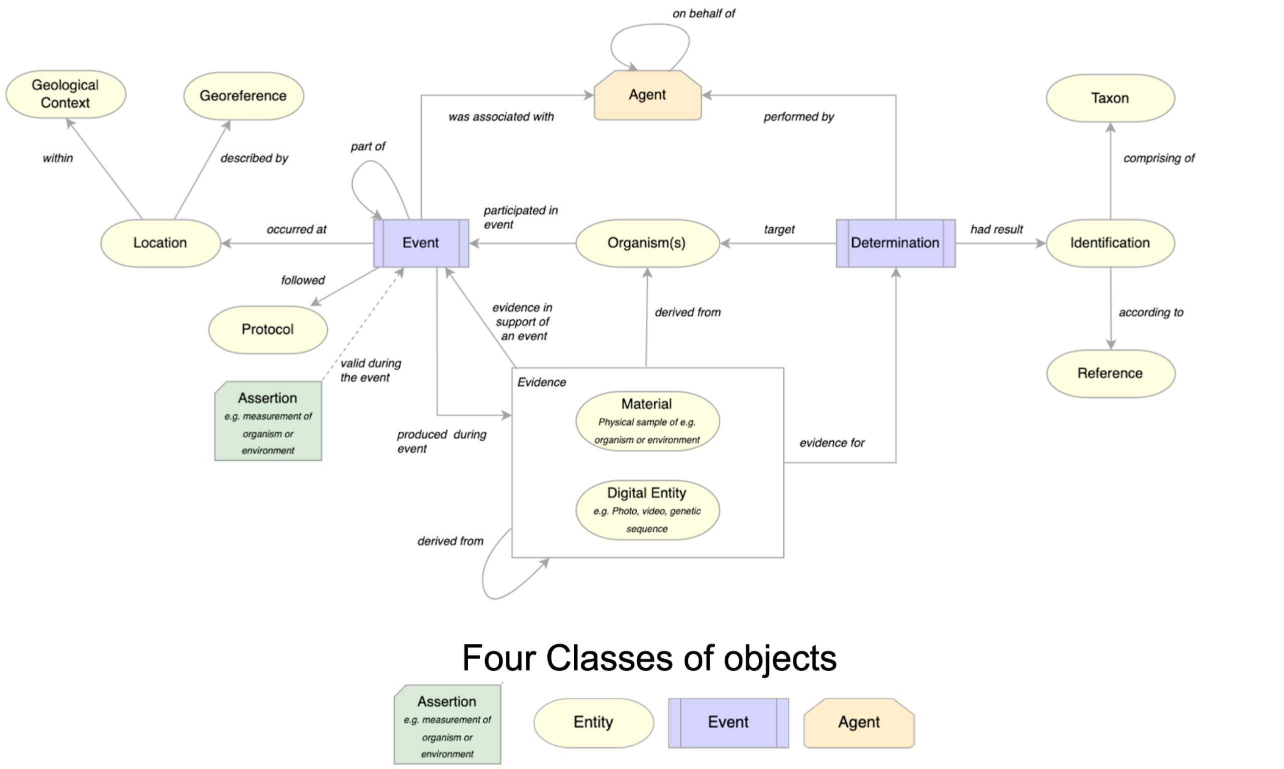


*Figure 4: Overview of some key concepts and their semantic linkages that cover many of the typical datasets shared through GBIF and form the basis of the Unified Model; it contains four foundational elements also found in Semantic Darwin Core.*

## 4. Example of application of the conceptual model

Figure 5 describes the application of the conceptual model with the component entities, events, agents and assertions. This example is built around a common GBIF data type, the preserved



'Global cooperation on FAIR data policy and practice' (WorldFAIR) has received funding from the European Union's Horizon Europe project call HORIZON-WIDERA-2021-ERA-01-01, grant agreement 101058393.

specimen, with several additional pieces of rich information, such as images, DNA sequences, and a new species description. The diagram, and therefore the data, can be described by the relationships of the foundational classes.

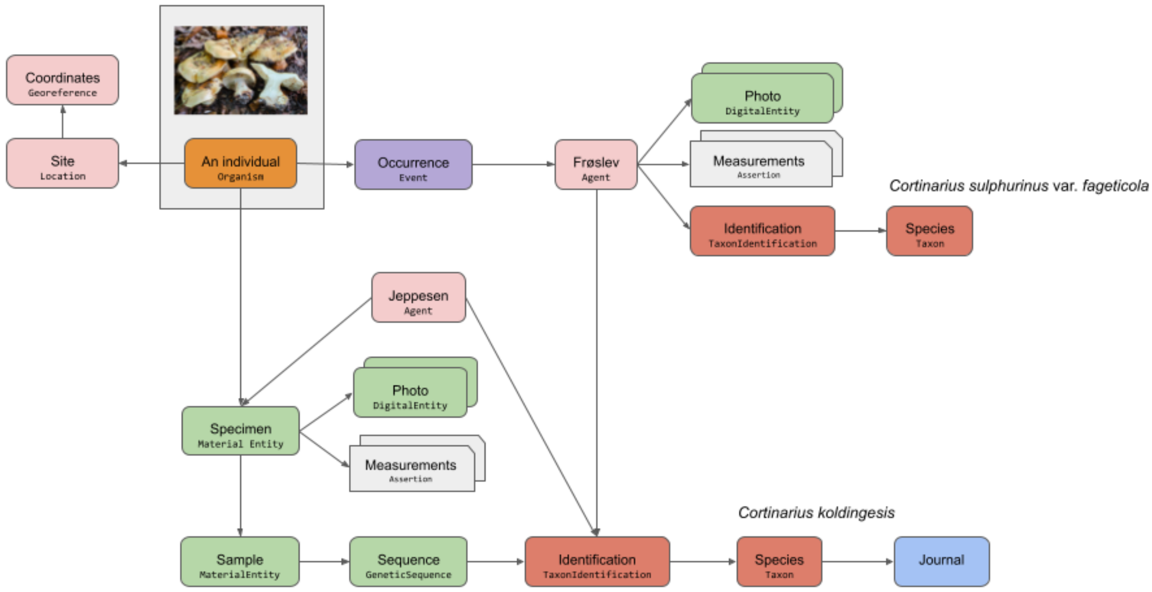


*Figure 5: An example of the conceptual model depicting the relationships around fungal data that would be typical to GBIF. Colours are not indicative.*

- Frøslev (Agent), observed the fruit of an Organism (Entity) at a site location (Entity) that was georeferenced. This represents the organism occurrence (Event) in nature. Frøslev (Agent) took digital photos (Digital Entity), made physical measurements (Assertions) and made an identification (Assertion) to a species (Entity).
- Frøslev (Agent) collected (Event) a fruit body sample (Material Entity) of the organism for preservation as a specimen. The specimen (Material Entity) was subsampled (Material Entity) and used as a source for DNA extraction which was sequenced (Digital Entity).
- Jeppesen (Agent) took more digital photos (Digital Entity) and physical measurements (Assertions) of the specimen.



'Global cooperation on FAIR data policy and practice' (WorldFAIR) has received funding from the European Union's Horizon Europe project call HORIZON-WIDERA-2021-ERA-01-01, grant agreement 101058393.

- Jeppesen, Frøslev (Agents) described a new species that was published in a journal (Entity) and used the DNA sequence (Digital Entity) and photos (Digital Entities) as evidence for the species being distinct from other known species.

- Many other Agents could have developed Events to create new Entities leading to new Assertions. All this and more is theoretically possible to layer on top of this conceptual model.

## 5. Test Use Case: Exploring collection management systems

Specimen data, such as items stored in natural history museums, was the foundational data type at GBIF's inception and remains the highest quality data type due to specimens stored as long-term evidence. The maturity of this data type and the sharing of data to GBIF by thousands of natural history museums make these material samples an ideal initial test case for the Unified Model. Another advantage of exploring natural history specimens and their materials is that there are many different types of collection management systems employed at natural history museums. Therefore this use case can establish if the model is robust and applicable to provide needed interoperability. GBIF held the webinar Exploring Collection Management Systems on this topic.

In this use case, the term 'material sample' is used as a general term to denote a biodiversity specimen stored in a natural history museum. In other parts of the Unified Model, a material sample can signify something other than a natural history specimen, such as a soil sample; therefore, the concept of a material sample can be employed in several aspects of the Unified Model.

The 'exploring collection management systems' example draws on two use cases: material sample and digital media data.

### 5.1 Material Sample

The preserved specimen is an integral part of this model and the model can reference it as well as any subsample of material from it that was taken for specific analysis such as DNA sequencing. The model is required to enable the discovery of specimens likely to have tissue still viable for DNA extraction. This use case affects biological material collections, botanical gardens, zoos, and gene banks as data (and potential material sample) providers.



'Global cooperation on FAIR data policy and practice' (WorldFAIR) has received funding from the European Union's Horizon Europe project call HORIZON-WIDERA-2021-ERA-01-01, grant agreement 101058393.

While this material specimen example is of a previously-living organism, the material specimen conceptual model can be applied to soil samples and even non-living specimens such as geological samples. These concepts are at an early stage of development.

The protocols used to preserve material can be of critical importance in determining its viability for a purpose. It is optimal for the model to expose the preservation methods used, so that the viability of a particular experiment is practical given the tissue. Though the use case is specific to tissues, the same concerns may apply to the full range of loanable material.

This use case (see Figure 6) requires common filtering capabilities for PreservedSpecimens, LivingSpecimens, FossilSpecimens (ancient DNA), and MaterialSamples. Key components of the Unified Model that are covered in this use case include:

- material preservation history that can separate all the states that a particular part of a specimen has been in through time. This includes the discoverability of already existing DNA extracts as stored material.
- identification history along with the list of specific materials used for each identification.

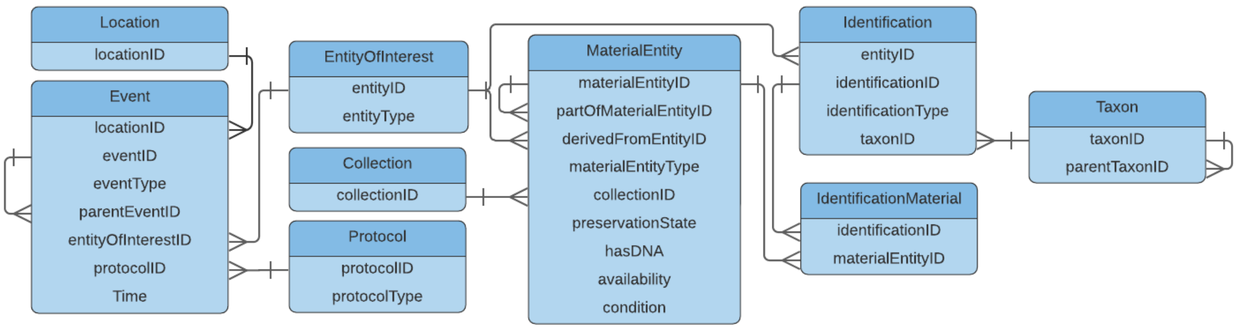


*Figure 6: A conceptual model covering the material samples from specimens.*

A collecting **Event** happens at a **Location** during a period of time (included in the **Event**), and may follow a specified **Protocol** with respect to the target **Organism** (an **EntityOfInterest**). The whole **Organism** or parts from it (**MaterialEntities**) may be gathered and optionally added to a **Collection**. The **MaterialEntities** may go through changes in preservation and availability, which can be tracked as additional **Events**. Each change-of-state **Event** spawns a new **MaterialEntity** record that is either part of or derived from an earlier **MaterialEntity** state. Multiple **Taxon Identifications** of the



'Global cooperation on FAIR data policy and practice' (WorldFAIR) has received funding from the European Union's Horizon Europe project call HORIZON-WIDERA-2021-ERA-01-01, grant agreement 101058393.

**Organism** can be made, each one using a potentially distinct set of **MaterialEntities** derived from the **Organism**.

## 5.2 Digital media data

DigitalEntities can also be generated from natural history specimens, that is to say the Organism and derivative MaterialEntities. Application of this part of the Unified Model requires the usual filtering capabilities for PreservedSpecimens, LivingSpecimens, FossilSpecimens, and MaterialSamples plus the presence of associated media.

Key components of the Unified Model covered in this use case include the following:

- accommodate media captured at distinct times and places in the history of an Organism and its derivative Material.
- accommodate associated media for which the specimen is not a subject, such as a photo of the place where the specimen was collected.
- accommodate MaterialEntities derived from DigitalMedia (a 3D-printed model).
- make identifications based on Media; see Use Case: Global Malaise Programme.

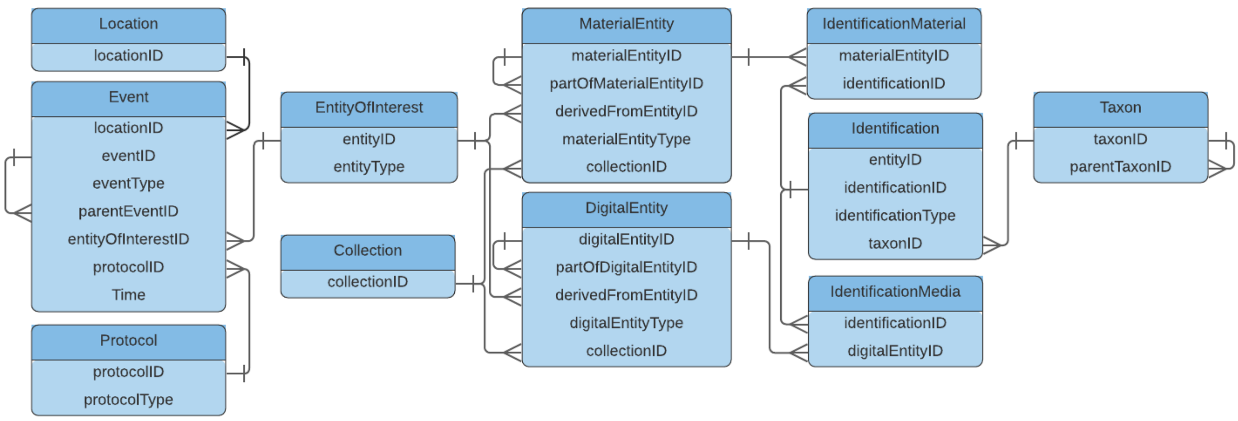


*Figure 6: A conceptual model covering the derivation of digital media from specimens.*

A collecting **Event** happens at a **Location** during a period of time (included in the **Event**), and may follow a specified **Protocol** with respect to the target Organism (an **EntityOfInterest**). The whole Organism or parts of it (**MaterialEntities**) may be gathered and optionally added to a **Collection**. Similarly, one or more **DigitalEntities** can be made of the Organism or **MaterialEntities** that are extracted from or form parts of it, the acts of which can be tracked as additional **Events**. Each **DigitalEntity** can also be added to a **Collection**.



'Global cooperation on FAIR data policy and practice' (WorldFAIR) has received funding from the European Union's Horizon Europe project call HORIZON-WIDERA-2021-ERA-01-01, grant agreement 101058393.

A **MaterialEntity** may be *partOf* another **MaterialEntity** (in situ) or *derivedFrom* another **MaterialEntity**. The *derivedFrom* and *partOf* relations should point only to the direct parent **MaterialEntity**. It might be useful in the physical model to include a key to the *parentmostEntityOfInterest* to shortcut to the highest source in the hierarchy, especially if that **Entity** is an Organism.

## 5.3 Community review of Collection Management Systems

In order to understand how the data can be published and then formatted via the Unified Model, GBIF commissioned twelve community use cases with users of various Collection Management Systems (CMS). These ongoing case studies are helping GBIF develop the appropriate data publishing model for the breadth of data available in the CMS. They also provide an opportunity to get community feedback on draft visualisation of collection data exposed via the new model. Below is a description of the use case developed for WorldFAIR (see Figure 7).



'Global cooperation on FAIR data policy and practice' (WorldFAIR) has received funding from the European Union's Horizon Europe project call HORIZON-WIDERA-2021-ERA-01-01, grant agreement 101058393.

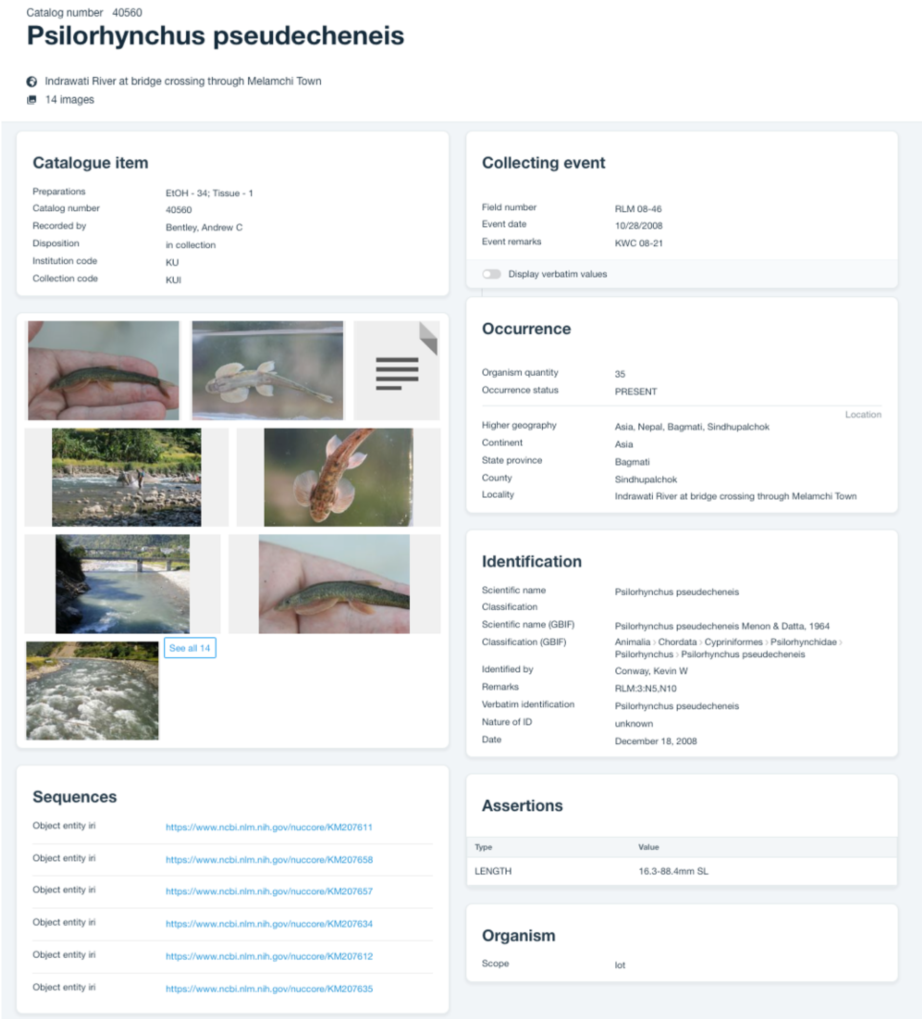


*Figure 7: Draft visualisation of a specimen modelled via the new Unified Model. This example describes the collection event, the catalogue item, occurrence, identification history, measurements (assertions) and DNA sequences. Compare this to the same preserved specimen in GBIF using current DwC star schema.*

## 6. WorldFAIR Use Case: Ecological monitoring with camera trap data

Development and improvement of various types of long-term ecological monitoring are two critical needs. This is evident from the Global Biodiversity Framework that came out from the Kunming-Montreal Global Biodiversity Framework COP 15 in December 2022. The European Green Deal and European Commission understands these needs and long-term monitoring is an essential component of many European funded activities including Biodiversa+, EuropaBon, Knowledge Centre for Biodiversity, and the European Environment Agency.



'Global cooperation on FAIR data policy and practice' (WorldFAIR) has received funding from the European Union's Horizon Europe project call HORIZON-WIDERA-2021-ERA-01-01, grant agreement 101058393.

Camera trap technology is growing in importance and will play a major role in biodiversity monitoring. Camera traps are a good example of a non-invasive technology to monitor entities, resulting in data, but current standards limit the ability to share data as required for policy impact.

A camera trap deployed in an area takes an image when triggered by movement. These images are the evidence used to detect an entity during the event. Multiple camera traps are often deployed to monitor large areas in concert.  Humans or artificial intelligence are used to make preliminary assertions about image contents (identification of individual animals, their sex, life stage, and taxonomic identification), and these are further vetted and improved by human experts.

Camera Trap Data Package (or Camtrap DP for short) is a community developed data exchange format for camera trap data. Camtrap DP is a Frictionless Data Package that consists of metadata, deployment table, media files and observation tables. The Camtrap DP model has been integrated in the Unified Model framework. This integration allows: 1) evidence that is based on the image, 2) individual organisms can be tracked between detections, 3) integration with artificial intelligence identification assertions, and 4) human assertions (see Figures 8, 9).

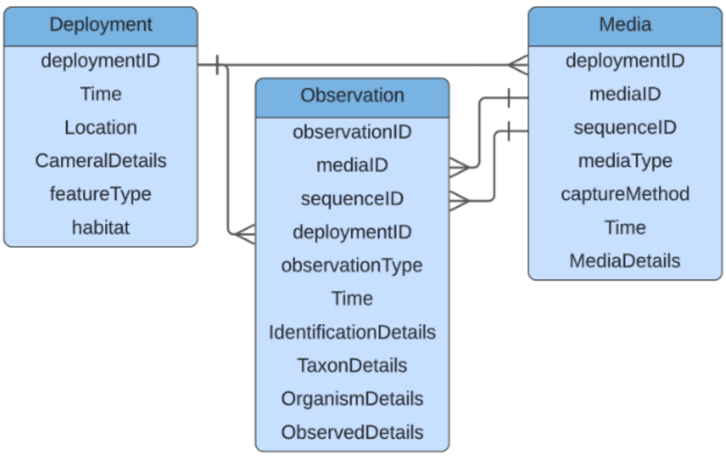


*Figure 8: A simplified version of the Camera Trap DP data model intended to be captured as a Frictionless Data Package. This closely resembles the [recommended publishing model](recommended publishing model).*



'Global cooperation on FAIR data policy and practice' (WorldFAIR) has received funding from the European Union's Horizon Europe project call HORIZON-WIDERA-2021-ERA-01-01, grant agreement 101058393.

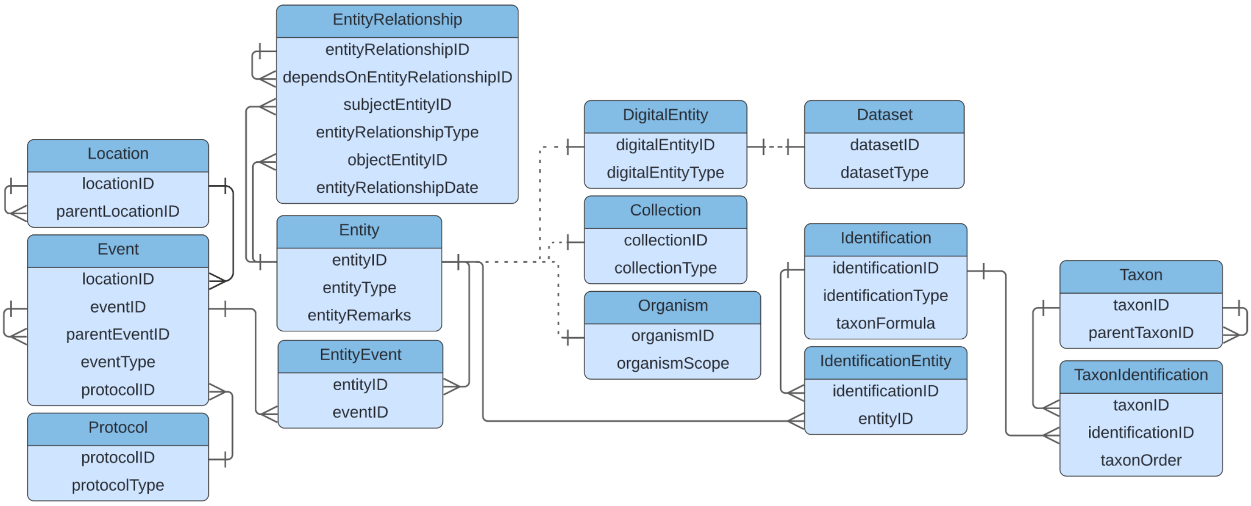


*Figure 9: A conceptual model covering the activities associated with a camera trap deployment, in which sequences result from detection triggers. Organism Identification and Occurrence information are interpreted from images by artificial intelligence routines and vetted by human group consensus.*

A deployment **Event** spans the period of time a camera is functioning properly in the field at a particular **Location** and may follow a specified **Protocol**. A sequence **Event** is triggered at the same **Location** as the deployment covering the field of view of the camera and spans a specific time period defining the bounds of the **Event**. Each sequence **Event** consists of one or more **StillImages** (each a **DigitalEntity**). An **Organism** (*sensu* Darwin Core; also an **Entity**) may be identified (taxonomically and potentially as a recurring individual) within an image **sequence** or within one or more **StillImages.** These **Organism** observations infer a **Taxon** occurrence (an observation **Event**) at the same time and place as the parent sequence or image **Event**. The observed **Organisms** can have **Assertions** about the count of individuals, life stage, sex, behaviour, etc. Each **Organism** can be given one or more **Taxon Identifications** by expert groups and/or by artificial intelligence processes. 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**.

The camera trap data model is under active development and was detailed in the webinar, [Exploring camera trap data](). In order to promote uptake, GBIF commissioned a guide to publishing camera trap data. This work includes testing the new Camtrap DP publishing model to GBIF infrastructure. This testing is in progress and will be released to community review in May with the final product made available later in 2023.


'Global cooperation on FAIR data policy and practice' (WorldFAIR) has received funding from the European Union's Horizon Europe project call HORIZON-WIDERA-2021-ERA-01-01, grant agreement 101058393.

## 7. Combining data from different sources

A great advantage of the Unified Model is that it permits many data publishing entry points and eventual delivery exits for data use as well. The following highlights a simple example from one of the twelve community use cases with Collection Management Systems. In this example a *Cyperus* herbarium specimen was collected in the late 19th Century in Togo and is stored and digitised in Berlin. As with most collections made at that time, the data is limited but the record contains an image. Recently a morphometric study used this very specimen (Xanthos et al. 2023) and supplied the detailed plant morphology measurements in the supplementary data of the publication. This is an example of the extended specimen – additional data that extends the value of the specimen (see Figures 10, 11).

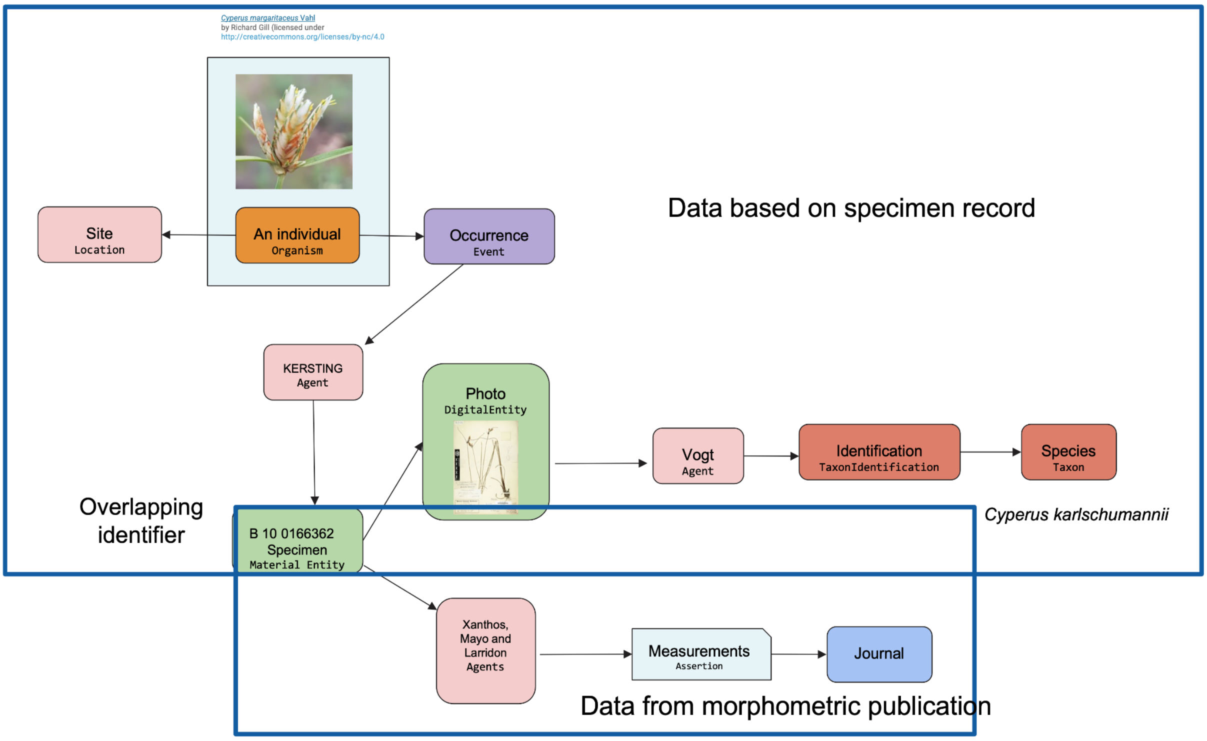


*Figure 10: An example of the conceptual model integrating data from two sources. The large box contains data from the specimen itself while the smaller box contains data that was published separately in a supplementary file in a publication. Normally these supplementary data are not connected to the specimen in a FAIR data repository.*



'Global cooperation on FAIR data policy and practice' (WorldFAIR) has received funding from the European Union's Horizon Europe project call HORIZON-WIDERA-2021-ERA-01-01, grant agreement 101058393.

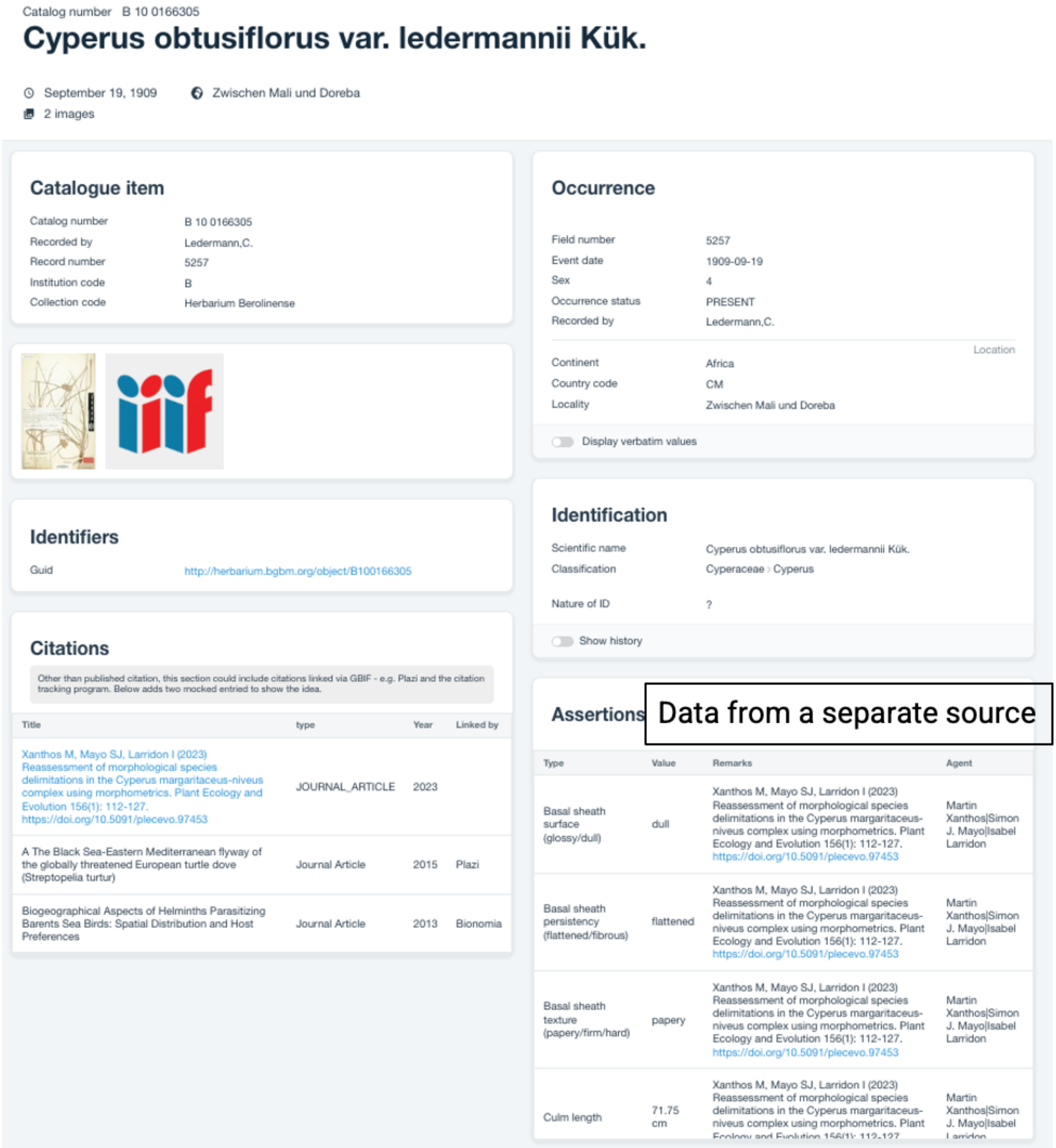


*Figure 11: Draft visualisation of a specimen modelled via the new Unified Model with data from two sources. This example describes the collection event, the catalogue item, occurrence, and identification history from the specimen voucher and the measurements (assertions) from a journal publication.*



'Global cooperation on FAIR data policy and practice' (WorldFAIR) has received funding from the European Union's Horizon Europe project call HORIZON-WIDERA-2021-ERA-01-01, grant agreement 101058393.

## 8. Status and next steps

During the first year of WorldFAIR, WP09's technical lead performed the tasks outlined in the proposal. First, a background analysis was undertaken to understand the history of DwC and related standards to identify models that could be adapted to enhance DwC. This work is detailed in Section 3 of this report, 'The Event: Initial observations on the Unified Model'. This work was successful in part because the lead in this WorldFAIR WP is a long-time lead contributor to the Darwin Core maintenance group, a committee of the Biodiversity Information Standards group, TDWG.

This integrated engagement with TDWG and Darwin Core facilitated professionalisation of the work which is otherwise a voluntary activity by community members. The project is highlighted on the Diversifying the GBIF Data Model page on the GBIF portal with links to online documents describing the case studies, which are open for community input. Once the Unified Model took shape, the team led three separate webinars on the following topics: 1) Diversifying the GBIF data model, 2) Exploring collection management systems, and 3) Exploring camera-trap data.

Testing and community engagement over the next year will continue to improve our knowledge and the model so that it better reflects the complexity of biodiversity data, leading to better use in research and policy. The feedback we have received also indicates a steep learning curve for the future implementation of the data model. This feedback is essential for the development of publishing tools.

This deliverable aligns well with the overall objectives of WorldFAIR by offering a detailed examination of how to improve the 'I' in FAIR, Interoperability, for biodiversity data. This work promotes cross-domain interaction, as the Unified Model will enhance sharing of data in related Work Packages such as Agricultural Biodiversity, Oceans and Geochemistry in the final portion of the grant period. This work has been undertaken in alignment with the overall WorldFAIR goals, in particular WP02 on Engagement, Synthesis, Recommendations and FAIR Assessment.

The development of a new Unified Model has had a significant and successful start in a long, multi-year process. The model foundations are set and the concerted development of case studies through open community involvement will continue to advance which will bring in more potential users and good ideas.

Currently we are developing publishing methods to implement the conceptual models. These publishing models can be simpler than the conceptual models and are critical to the uptake and eventual success of the Unified Model. This part of the process will perhaps be the most important


'Global cooperation on FAIR data policy and practice' (WorldFAIR) has received funding from the European Union's Horizon Europe project call HORIZON-WIDERA-2021-ERA-01-01, grant agreement 101058393.

and time-consuming.  Once data is available in the Unified Model, there will be many options for innovative data use. We foresee new data catalogues for particular user communities. As an infrastructure and a network of people and organisations, GBIF will heavily rely on guidance in pursuing these opportunities.

## 9. Acknowledgements

Like GBIF itself, the development of this Unified Model is a community endeavour. GBIF is taking the lead to develop a model that will improve GBIF mediated data and we expect it will improve data in the larger community as well. Funding for this project comes from the GBIF core budget, the WorldFAIR grant and collaboration with the larger community, much of it voluntary. This includes the Biodiversity Information Standards (TDWG) and the Atlas of Living Australia.



'Global cooperation on FAIR data policy and practice' (WorldFAIR) has received funding from the European Union's Horizon Europe project call HORIZON-WIDERA-2021-ERA-01-01, grant agreement 101058393.

# 10. Bibliography

Baskauf, Steven & Webb, Campbell. (2016). Darwin-SW: Darwin Core-based terms for expressing biodiversity data as RDF. Semantic Web. 7. 629-643. 10.3233/SW-150203.

Baskauf, Steve and Joel Sachs. (2018). "Darwin Core as a Vocabulary for Expressing Biodiversity Data as RDF." Application of Semantic Technology in Biodiversity Science.

Casati, Roberto & Varzi, Achille C. (2020). Events. Stanford Encyclopedia of Philosophy.

CODATA, the Committee on Data of the International Science Council, Pfeiffenberger, Hans, Uhlir, Paul, & Hodson, Simon. (2020). Twenty-Year Review of GBIF. https://doi.org/10.35035/ctzm-hz97

Cox, S. (ed.) (2011) Observations and Measurements - XML Implementation, Version 2.0. Wayland, MA, Open Geospatial Consortium, 77pp. (OGC 10-025r1). DOI: http://dx.doi.org/10.25607/OBP-627

GBIF Strategic Framework 2023-2027 (2021) GBIF Secretariat: Copenhagen. https://doi.org/10.35035/doc-0kkq-0t82

Haller, Armin & Janowicz, Krzysztof & Cox, Simon & Phuoc, Danh & Taylor, Kerry & Lefrançois, Maxime. (2017). Semantic Sensor Network Ontology.

Heberling JM, Miller JT, Noesgaard D, Weingart SB and Schigel D (2021) Data integration enables global biodiversity synthesis. Proceedings of the National Academy of Sciences. Proceedings of the National Academy of Sciences 118(6): e2018093118. Available at: https://doi.org/10.1073/pnas.2018093118.

Lebo, T, Sahoo, S., McGuinness D. eds. PROV-O: The PROV Ontology. 30 April 2013, W3C Recommendation. URL: http://www.w3.org/TR/2013/REC-prov-o-20130430/

Niles, I, and Pease, A . 2001. Towards a standard upper ontology. In Proceedings of the international conference on Formal Ontology in Information Systems - Volume 2001 (FOIS '01). Association for Computing Machinery, New York, NY, USA, 2–9. https://doi.org/10.1145/505168.505170

Sowa, J. (2000). Knowledge Representation: Logical, Philosophical, and Computational Foundations.

Wieczorek J, Bloom D, Guralnick R, Blum S, Döring M, Giovanni R, et al. (2012) Darwin Core: An Evolving Community-Developed Biodiversity Data Standard. PLoS ONE 7(1): e29715. https://doi.org/10.1371/journal.pone.0029715

Walls, Ramona & Deck, John & Robert, Guralnick & Baskauf, Steve & Beaman, Reed & Blum, Stan & Bowers, Shawn & Buttigieg, Pier Luigi & Davies, Neil & Endresen, Dag & Gandolfo, Maria & Hanner, Robert & Janning, Alyssa & Krishtalka, Leonard & Matsunaga, Andréa & Midford, Peter & Morrison, Norman & Tuama, Eamonn & Schildhauer, Mark & Wooley, John. (2014). Semantics in Support of Biodiversity Knowledge Discovery: An Introduction to the Biological Collections Ontology and Related Ontologies. PloS one. 9. e89606. 10.1371/journal.pone.0089606.



'Global cooperation on FAIR data policy and practice' (WorldFAIR) has received funding from the European Union's Horizon Europe project call HORIZON-WIDERA-2021-ERA-01-01, grant agreement 101058393.

Xanthos M, Mayo SJ, Larridon I (2023) Reassessment of morphological species delimitations in the *Cyperus margaritaceus-niveus* complex using morphometrics. Plant Ecology and Evolution 156(1): 112-127. https://doi.org/10.5091/plecevo.97453



'Global cooperation on FAIR data policy and practice' (WorldFAIR) has received funding from the European Union's Horizon Europe project call HORIZON-WIDERA-2021-ERA-01-01, grant agreement 101058393.