Robust Integration of Biodiversity Data by Process- and State-based Representation of Object Histories and Modular Application Architecture

Christian Bölling[‡], Satpal Bilkhu[§], Christian Gendreau[§], Falko Glöckler[‡], James Macklin[§], David Shorthouse[§]

‡ Museum für Naturkunde Berlin, Leibniz Institute for Evolution and Biodiversity Science, Berlin, Germany
§ Agriculture and Agri-Food Canada, Ottawa, Canada

Corresponding author: Christian Bölling (christian.boelling@mfn.berlin)

Abstract

Biodiversity data is obtained by a variety of methodological approaches—including observation surveys, environmental sampling and biological object collection—employing diverse sample processing protocols and data transformations. While complete and accurate accounts of these data-generating processes are important to enable integration and informed reuse of data, the structure and content of published biodiversity data currently are often shaped by specific application goals. For example, data publishers that export specimen-based data from collection management systems for inclusion in aggregations like those in the Global Biodiversity Information Facility (GBIF) must frequently relax their internal models and produce unnatural joins to fit GBIF's occurrences-based data structure. Third-party assertions over these aggregated data therefore assume the risk of irreproducibility or concept drift.

Here we introduce process- and state-based representation of object histories as the main organizing principle for data about specimens and samples in *Digital Information System for Natural History Data* (DINA, Glöckler et al. 2020)-compliant collection management software (Fig. 1). Specimens, samples and objects in general are subjected to a variety of processes, including planned actions involving the object, e.g., collecting, preparing, subsampling, loaning. Object states are any particular mode of being of an object at a certain point in time. For example, any one intermediate step in preparing a collected specimen for long-term conservation in a collection would constitute an individual object state. An object's history is the entire chain of these interrelated processes and states.

We argue that using object histories as main conceptual modeling paradigm in DINA offers the generality required to accommodate a diverse, open set of use cases in biodiversity data representation, yet also offers the versatility to serve as basis for use-case specific

data aggregation and presentation. Specifically, a representation based on object histories provides

- a coherent structure for documenting individual processes and states for any given object and for linking this documentation (e.g., textual descriptions or images pertaining to a given process or state),
- a natural representational structure of the real-world sequence of processes an object participates in and for the data generated in these processes (e.g., a DNA-extraction procedure and sequence information generated on its basis),
- a straightforward structure to link data about related objects (e.g., tissue samples, the biological specimen a bone is derived from) in a network of connected object histories.

The approach is designed to be embedded in DINA's modular application architecture, so that information on object histories can be accessed via corresponding APIs either through its own interfaces (Fig. 2) or by integration with external web services (Fig. 3). Viewing collection management tasks as part of object histories also informs delineation of modules to support these tasks with specialized functions and interfaces. It also admits the use of persistent, dereferencable identifiers for individual processes and states in object histories and for linking their representations to elements in ontologies and controlled vocabularies.

In this contribution to the symposium, DINA's object histories as a main organizing principle for collection object data will be discussed and the utility of using it in the context of modular application architecture, data federation, and data integration in projects like BiCIKL will be illustrated.

Keywords

collection management, biodiversity knowledge graph, DINA

Presenting author

Christian Bölling

Presented at

TDWG 2021

Conflicts of interest

References

- Glöckler F, Macklin J, Shorthouse D, Bölling C, Bilkhu S, Gendreau C (2020) DINA—Development of open source and open services for natural history collections & research. Biodiversity Information Science and Standards 4 https://doi.org/10.3897/biss.4.59070

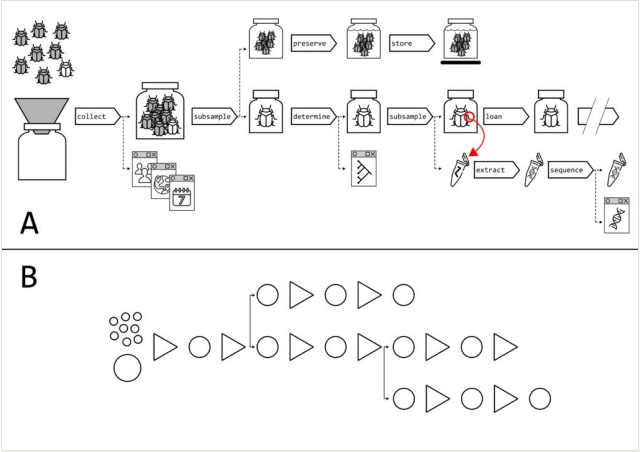

Figure 1.

(A) Concept schema for interconnected object histories (example). (B) Corresponding abstract concept schema of sequences of processes (triangles) and objects in particular states (circles) before and after processes. Processes give rise to new objects, object states and data items about objects and processes. Processes in the context of collection management will usually consist of separable sub-processes (e.g., different steps in a preservation protocol). Objects can be an aggregate of other objects (e.g., a lot of biological specimens and the preparation container as a composite object originating from a preparation process).

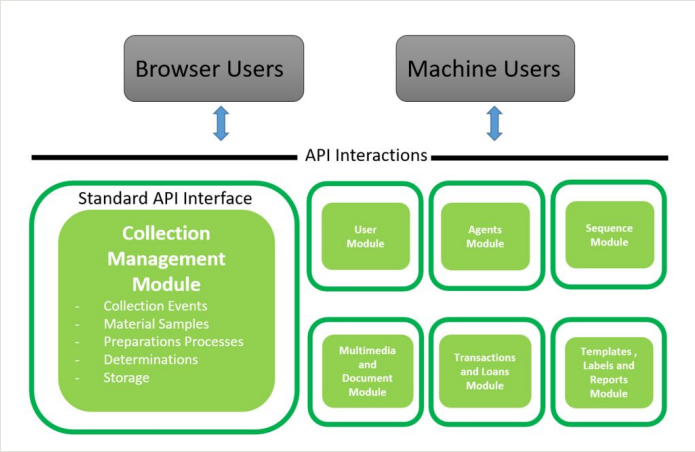

Figure 2.

In DINA's modular software, architecture modules communicate with other modules through standard application programming interfaces (APIs) as do browser or machine users when interacting with the system.

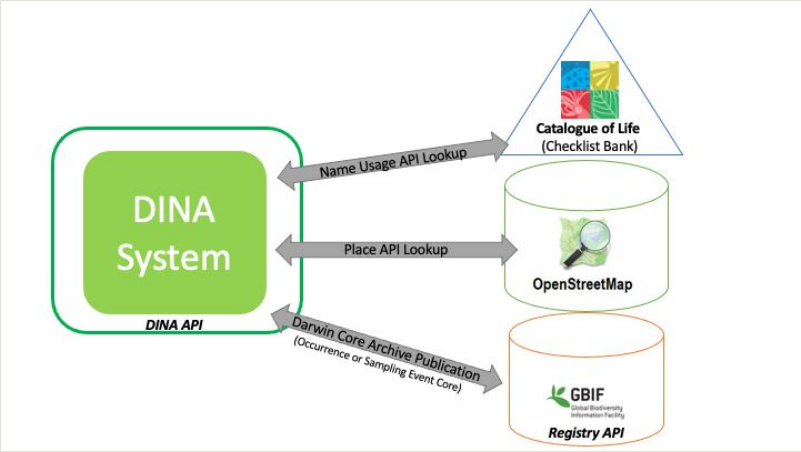

Figure 3.

Application programming integration of the DINA system with the Catalogue of Life (Checklist Bank) to anchor scientific names in determinations, Open Street Maps to anchor place names in georeferences, and with the GBIF registry to publish occurrences and sampling event core data sets.