BISS Biodiversity Information Science and Standards

Biodiversity Information Science and Standards 3: e37422
doi: 10.3897/biss.3.37422

OPEN ACCESS

Conference Abstract

Managing Taxon Data in FinBIF

Esko Piirainen‡, Eija-Leena Laiho‡, Tea von Bonsdorff‡, Tapani Lahti‡

‡ Finnish Museum of Natural History LUOMUS, Helsinki, Finland

Corresponding author: Esko Piirainen (esko.piirainen@helsinki.fi)

Received: 17 Jun 2019 | Published: 26 Jun 2019

Citation: Piirainen E, Laiho E, von Bonsdorff T, Lahti T (2019) Managing Taxon Data in FinBIF. Biodiversity Information Science and Standards 3: e37422. https://doi.org/10.3897/biss.3.37422

Abstract

The Finnish Biodiversity Information Facility, FinBIF (https://species.fi), has developed its own taxon database. This allows FinBIF taxon specialists to maintain their own, expert-validated view of Finnish species. The database covers national needs and can be rapidly expanded by our own development team. Furthermore, in the database each taxon is given a globally unique persistent URI identifier (https://www.w3.org/TR/uri-clarification), which refers to the taxon concept, not just to the name. The identifier doesn't change if the taxon concept doesn't change. We aim to ensure compatibility with checklists from other countries by linking taxon concepts as Linked Data (https://www.w3.org/wiki/LinkedData) — a work started as a part of the Nordic e-Infrastructure Collaboration (NeIC) DeepDive project (https://neic.no/deepdive).

The database is used as a basis for observation/specimen searches, e-Learning and identification tools, and it is browsable by users of the FinBIF portal. The data is accessible to everyone under CC-BY 4.0 license (https://creativecommons.org/licenses/by/4.0) in machine readable formats.

The taxon specialists maintain the taxon data using a web application. Currently, there are 60 specialists. All changes made to the data go live every night. The nightly update interval allows the specialists a grace period to make their changes. Allowing the taxon specialists to modify the taxonomy database themselves leads to some challenges. To maintain the integrity of critical data, such as lists of protected species, we have had to limit what the specialists can do. Changes to critical data is carried out by an administrator.

The database has special features for linking observations to the taxonomy. These include hidden species aggregates and tools to override how a certain name used in observations is linked to the taxonomy. Misapplied names remain an unresolved problem. The most precise way to record an observation is to use a taxon concept: Most observations are still recorded using plain names, but it is possible for the observer to pick a concept. Also, when data is published in FinBIF from other information systems, the data providers can link their observations to the concepts using the identifiers of concepts. The ability to use taxon concepts as basis of observations means we have to maintain the concepts over time — a task that may become arduous in the future (Fig. 1).

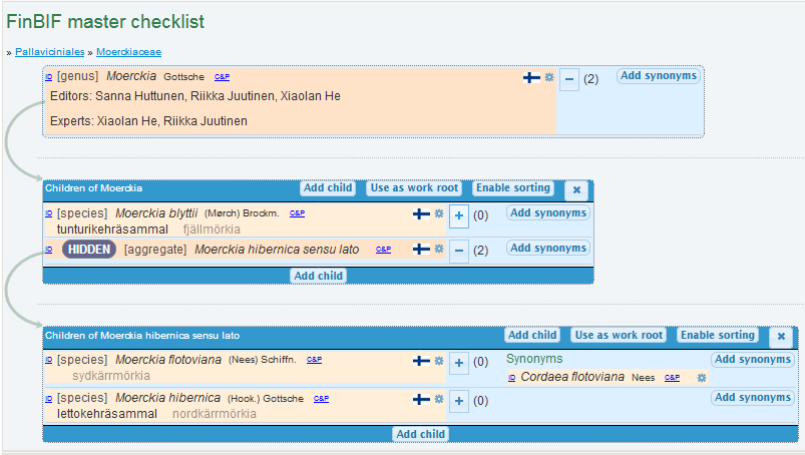

Figure 1.

A screenshot of FinBIF Taxon Editor web application. The figure illustrates a scenario where a taxon concept has been split into two concepts. The original concept remains in the database as a hidden aggregate.

As it stands now, the FinBIF taxon data model — including adjacent classes such as publication, person, image, and endangerment assessments — consists of 260 properties. If the data model were stored in a normalized relational database, there would be approximately 56 tables, which could be difficult to maintain. Keeping track of a complete history of data is difficult in relational databases. Alternatively, we could use document storage to store taxon data. However, there are some difficulties associated with document storages: (1) much work is required to implement a system that does small atomic update operations; (2) batch updates modifying multiple documents usually require writing a script; and (3) they are not ideal for doing searches. We use a document storage for observation data, however, because they are well suited for storing large quantities of complex records.

In FinBIF, we have decided to use a triplestore for all small datasets, such as taxon data. More specifically, the data is stored according to the RDF specification (https://www.w3.org/RDF). An RDF Schema defines the allowed properties for each class. Our triplestore implementation is an Oracle relational database with two tables (resource and statement), which gives us the ability to do SQL queries and updates. Doing small atomic

updates is easy as only a small subset of the triplets can be updated instead of the entire data entity. Maintaining a complete record of history comes without much effort, as it can be done on an individual triplet level. For performance-critical queries, the taxon data is loaded into an Elasticsearch (https://www.elastic.co) search engine.

Keywords

taxonomy, taxon concepts, linked open data, observations, database solutions, triplestore, ontology, RDF, Elasticsearch, FinBIF

Presenting author

Esko Piirainen

Presented at

Biodiversity_Next 2019