An Algorithmic Approach to Reducing Taxonomic Detail from Actual Datasets to their Metadata Representation to Increase Findabilty

Cedric Decruw[‡], Leen Vandepitte[§], Ruben Perez Perez[‡], Laura Marquez[‡], Marc Portier[‡], Bart Vanhoorne[‡], Lennert Tyberghein[‡]

‡ Flanders Marine Institute, Ostend, Belgium
§ Vlaams Instituut voor de Zee (Flanders Marine Institute) - VLIZ, Oostende, Belgium

Corresponding author: Cedric Decruw (cedric.decruw@vliz.be)

Abstract

The rise in demand for more FAIR (Findable, Accessible, Interoperable, and Reusable) data is being answered by increasingly automated ways to capture, process, publish and register biodiversity datasets. Coupled with the increasing possibilities for detecting hundreds of species in a single sample/event (i.e., eDNA), this results in taxonomic information that is multiple levels of magnitude higher than it was a couple of years ago. This spike in content has an adverse effect on the ability of researchers to find relevant datasets within catalogues, due to the limitations in storing and displaying the taxonomic metadata (e.g., in the real-estate of a webpage, in the timeframe required to access the quantity of information, in displaying information to users in a comprehensive and comprehensible way).

WoRMS (World Register of Marine Species) is a taxonomic backbone that provides species information. One user of WoRMS is the Integrated Marine Information System (IMIS), a metadata catalogue for marine data, which is also the metadata catalogue of the European node of the Ocean Biodiversity Information System (EurOBIS). Taxonomic information added to the metadata records in IMIS are linked to WoRMS Persitent Identifiers or PIDs (AphiaIDs). Tension between providing all the taxonomic metadata while not overloading the catalogue is being addressed for the use-case of WoRMS+IMIS+EurOBIS.

Our approach is to apply a filter-and-replace algorithm during the automated registration of the taxonomic metadata to describe available datasets. This technique reduces the detailed taxonomic information of actual occurrences in the dataset content into practical (good enough) metadata. It takes as input all the species in the dataset along with their hierarchical structure, as well as a configuration parameter allowing for an upper bound to the acceptable number of taxa to be output.

The core principle of the algorithm is to start off with the minimal result set containing only the hierarchical root (always "biota") of the complete taxonomy in the dataset, and then to gradually consider replacing each element with its children one level deeper, as long as that replacement keeps fitting the upper bound for the total set.

This approach ensures that no coverage is lost, meaning every taxon in the actual dataset is represented in the result, although possibly through one of its parents, X layers up. Note that as long as only one child is underlying, the switch will always happen. So by nature, it will go down to the lowest relevant detail without challenging the upper bound limit.

It also allows for variation in the actual processing by allowing for different ordering strategies on the current result-set. Ordering strategies under test are:

- Order (descending) by weight of underlying available children → favouring more detail in those parts of the tree that have the most members in the set
    - With weight defined as the count of all underlying available species, OR
    - Weight defined as the sum-product of those species with their actual occurrences in the dataset (thus further favouring detail to those parts of the tree that are more prevalent in the samples)
- Order (descending) by number of direct children → favouring fanning-out over available high-level siblings
- Order (ascending) by ratio of present children over available children in the taxa → favouring replacing too vague parents with the more specific sublevels that are actually in the dataset

Alongside this basic approach, additional pre-processing of the taxa in the dataset can apply some form of "pruning". In this approach all nodes in the available taxon tree of the dataset are ordered (descending) by the weight of underlying children, and those at the end—below some defined cutoff ratio (extra parameter to the algorithm)—are simply discarded. Again, the interpretation of this weight (only species count, or multiplied by occurrence count) yields to variants of the algorithm to be tested. Applying this pruning means a deliberate departure is taken from the full-coverage guarantee mentioned earlier. Caution should be applied, of course, but removing more irrelevant (low occurrence) parts of the tree will allow for making space in the bounded result-set for more detail in the parts that are relevant.

In order to create an objective basis for comparing the resulting variants, a number of "qualification" parameters are considered to quantify the effect of the suggested reduction. Based on the datasets in EurOBIS, the variants of this algorithm are being applied and results will be presented on how they affect the various qualification parameters.

It is worth observing that both datasets and their metadata records are distinct resources that are being linked to species (taxa references) and that the different purposes they serve require different levels of detail to be presented. As other types of entities

(publications, habitats, experts, geography, traits, etc.) are considered for linking to species, we believe similar reduction algorithms will be necessary.

Keywords

taxonomy (biology), FAIR, algorithm, occurrence

Presenting author

Cedric Decruw

Presented at

TDWG 2022

Funding program

EMODNet Biology: This work has been financially supported by the EC DG-MARE (EMODnet Observation and Data network - Lot5 - Biology: EASME/EMFF/2016/1.3.1.2/ Lot5/SI2.750022).

Hosting institution

Flanders Marine Institute (VLIZ)

Conflicts of interest