Biodiversity Information Science and Standards 3: e35245
doi: 10.3897/biss.3.35245

Conference Abstract

Lessons Learned from Managing Millions of Records to Create the World Flora Online

William Ulate [‡]

[‡] Missouri Botanical Garden, St. Louis, MO, United States of America

Corresponding author: William Ulate (william.ulate@mobot.org)

Received: 07 Apr 2019 | Published: 13 Jun 2019

Citation: Ulate W (2019) Lessons Learned from Managing Millions of Records to Create the World Flora Online. Biodiversity Information Science and Standards 3: e35245. https://doi.org/10.3897/biss.3.35245

Abstract

In 2013, the World Flora Online (WFO) Consortium Council decided to use version 1.1 of the The Plant List (TPL) to initially populate the WFO taxonomic backbone. TPL is a collaboration between the Royal Botanic Gardens, Kew, Missouri Botanical Garden and other stakeholders to create a comprehensive list of Vascular plant (flowering plants, conifers, ferns and their allies) and of Bryophytes (mosses and liverworts). By combining multiple checklist held by these institutions, TPL 1.1 contained 1,064,035 scientific plant names of species rank, 350,699 of which were accepted species names. TPL provides the **Accepted** Latin names linked to **Synonyms** by which that species has been known. It also includes **Unresolved** names for which the contributing data sources did not contain sufficient evidence to decide whether they were **Accepted** or **Synonyms**.

Fortunately, TPL keeps track of the provenance of names and links back to the International Plant Names Index (IPNI) repository. This provenance trace has proven crucial when giving proper credit, as well as implementing a reliable curating process in WFO that supports the incorporation of potential new content, updates and further improvements contributed by the source. We will see some examples in WFO where duplication of names is originated from combining different providers and different sources, but also cases where duplication was caused within the same provider and even within a single source.

The WFO Council also decided to adopt the software used by eMonocot.org to display and harvest the information of plants. This decision made it possible to take advantage of the efforts previously done by the Monocots group in using already defined standards and existing tools to create and validate the input files harvested. Unfortunately, no technical documentation nor support was available for the eMonocot software and adapting the software code was not an option then. Therefore, a process of reverse engineering was implemented to determine what input was expected, which harvested values were actually stored in the database and what impact, if any, they could have on the Portal function. For example, the eMonocot software always harvests content under a particular hierarchy where an authority, in this case corresponding to a family taxon, holds ownership of the taxa underneath. We will explain how this may become an issue when incorporating new endemic taxa.

To ensure a convenient quality control, processes of validation and data curation were implemented. WFO assigns a unique ID to each name in its taxonomic backbone. The guarantee of uniqueness and permanence of such IDs is essential to support a process of cumulative improvement. To obtain this ID, a tool that matches Names was developed, allowing providers to contribute revisions to the taxonomy and descriptive content associated to a taxon. The origin of changes needs to be considered when tracing and correcting errors, implementing modifications or rolling back them later.

A report about the result of requested changes in the taxonomy needs to be approved by the provider before any actual change is implemented in the taxonomic backbone. Programmatically, any process that performs quality assessment or makes data modifications must be implemented as parameterized algorithms to allow replication of the process whenever new or updated data is available from the source. Single-use scripts are quick but not very scalable.

Finally, having defined a schema to use when providing content doesn’t necessarily imply that the values provided in each field are correct. Even with standardized values, the semantics associated could cause unforeseen behavior in the process implemented by the software. When possible, an additional step was required to convert harvested data from different localized vocabularies for standardized fields. Examples in Portuguese and Turkish will be given.

Keywords

Quality Control, Plants, Flora, WFO, World Flora Online, TPL, The Plant List, Taxonomic Backbone, Taxonomic Aggregators, Portal

Presenting author

William Ulate