[Edit Platform Documentation pt. xx]

Assembly of a preliminary taxonomic backbone for the Caryophyllales network's contribution to WFO

Version: 28-May-18 11:22

Contents

1 Workflow

This document describes import procedures for the EDIT Platform, in the context of the World Flora Online (WFO) contribution of the Caryophyllales Network. The BGBM is responsible for the taxonomic backbone of Caryophyllales in the World Flora Online. The aim is to establish a workflow to update the WFO Taxonomic Backbone (currently: data from a certain version of The Plant List) with data from the EDIT Platform treatments of Caryophyllales. Initially, this implies that Names and the corresponding WFO Name Identifier (and in some cases classifications) are imported from the existing WFO backbone. In the long run, only the updating of the WFO backbone will matter, whenever the Caryophyllale Network backbone

1.1 Prerequisites

- Operations concerning the WFO taxonomic backbone used in the WFO data portal are carried out by the WFO Data Center.
- An update of the WFO backbone from an external source (in our case: from a treatment in the EDIT Platform's Caryophyllales_spp database) implies the complete replacement of that specific part (normally a family) in the WFO backbone with the data from the external source.
- All names for that family that were already present in the WFO taxonomic backbone must be treated in the Platform (potentially only as "unresolved names" of different kinds). This is to acertain that all data linked into the WFO backbone (e.g. from Floras etc.) will continue to be linked to a name once the WFO backbone is updated (replaced).
- All data for that family in the Platform should eventually have a WFO ID, either the one already received from WFO or a new one taken from our contingent of WFO-IDs.

1.2 Basic process

- Obtain existing backbone name data with WFO-IDs from the WFO Data Center.
- Assemble a complete taxonomic backbone in the Platform where all names have WFO Id's and all existing WFO-backbone names are treated.
- Export the Platform data into a WFO DwC-A file and send that to the WFO Data Center.
- WFO Data Center checks the conformity to WFO DwC Standard and the presence of WFO-IDs.
- WFO Data Management replaces the WFO backbone for the respective taxonomic group with the data exported from the Platform.

1.3 Steps

- Selection of the taxonomic group – i.e. a single taxon node in the general Caryophylllaes backbone (normally a familiy)

- Requesting and receiving from the WFO Data Center a DwC-Archive containing the complete list of names and the respective WFO-IDs for that taxonomic group in that family's present WFO backbone. (See section 4 for the format).
- Pre-processing of the WFO data (see section
- Decisions:
    - If the taxonomic group has already (partially) been treated in the Platform and the treatment should be preserved[1] **Procedure A**
    - If the taxonomic group is not present in the Platform or can be replaced
        - Only Names should be imported **Procedure B**
        - The classification of the WFO-Backbone should be imported **Procedure C**
-

1.4 Procedure A: CDM-WFO Name Matching and WFO-ID Assignment

- Export of the names of the taxonomic group using the CDM-light export functionality of the TaxEditor
- Matching the Platform-names with the names in the table obtained from the WFO Data Center[2]
- Preparing a table containing the WFO-IDs and the corresponding Platform CDM-IDs
    - Format: See section 2.1
    - For those Names from the Platform that do not carry a WFO-ID: Assign new WFO-ID from Set of free IDs obtained from the WFO Data Centre
    - Optional: Inlude IPNI-IDs existing in the WFO dataset
- Import of that table (in effect: assigning WFO-IDs to the Names in the Platform, where possible)
    - [N.B.: Character encoding for Platform import: UTF-8]
- Exclude those names from the dataset obtained from the WFO Data Centre
- Decision on import options for the remaining data:
    - Only names and nomenclatural data should be imported **Procedure B**
    - The classification of the WFO-Backbone should be imported, too **Procedure C**

[1] An important criterium is the the amount of revision that went into nomenclatural citation and nomenclatural status, relationships, etc. of the names – classification changes (acceptance of taxa etc.) can rather easily be effected lateron in the Taxonomic Editor).

[2] This needs some manual work. Matching can only be done effectively between names without authors. But even so, the match often fails because of discrepancies in the generic gender assigned to the names and consequent differences in endings (-us, -a, etc.). WGB has a VBA function that reduces epithets to the stem, which works fairly well. (-us, -a, etc.) nicht. In a next step, those CDM names that did not match any in the WFO data need to be controlled manually for other orthographical variations. Moreover, the matching of names without authors results in wrong matches for homonyms – this needs to be looked at by revising the entire list manually.

1.5 Procedure B: Names-Only Import

Decision making requires an understanding of the treatment of names, taxa, synonyms, and classifications in the EDIT Platform, see section 5 for an explanation. A table of names with their name-related data (authors, nomenclatural citation, nomenclatural status, etc.) can be imported as names or as "pseudotaxa".

1.5.1 Importing Names as Names

The consequence of importing names as names is that they do not show up neither as taxon names nor synonyms in a classification, and consequently they remain invisible in the portal until something was done with them using the TaxEditor[3]. After import, they become available in the TaxEditor to be assigned to existing taxa as synonyms, or declared as a correct taxon name, or be related to another name in the context of a name relationship.

Steps:

- From the WFO-Data, create a table according to the specification given in section 2.3

1.5.2 Importing Names as "Pseudotaxa"

This imports the names and assignes them a (preliminary) taxonomic status as correct names for a taxon. This may have certain editorial advantages and disadvantages. One of the advantage is that the names become visible in the taxonomic tree and functionalities available there can be used to sort them out (e.g. the drag-and-drop assignment, right-click context menu). This also implies, that the editing person has a direct view on what names have not yet been treated properly. The disadvantage, especially in the case of large amounts of names is that these operations may get rather complicated or sluggish when carried out in the TaxEditor.

The pseudotaxa may be included in a separat classification or as a taxon-node in an existing classification. The former has the consequence that they can be excluded easily from a cdm-export and from showing up in the Platform Data Portal.

Steps:

- Decide, where to include the Pseudotaxa
- From the WFO-Data, create a table according to the specification given in section 2.3

[3] At present, it is not possible to list names without such usage in the TaxEditor.

1.6 Procedure C: Importing the WFO Backbone Classification

We have received a DwC-A export file containing all Caryophyllales names in the WFO Taxonomic Backbone including their taxonomic status (accepted, synonym [of which accepted name] and classification (assignment to higher taxon, i.e. family, genus, and species in case of infraspecific taxa).

The original file is structured according to the metadata given in section 4. We also have an Access database with the data. Here we can make some corrections to obvious errors. More importantly, we can assign the correct publication type (Article in periodical or Book) to the name, which is an important distinction in the import process.

2 Table definitions

2.1 DwC-A items accepted for import to a CDM database

Field name	Example	Destination
	6f44745f-dbc2-4dd6-8913-06d34568035d	
	wfo-0001302541	

2.2 Table of results from WFO-CDM Name Matching for Import into CDM

Field name	Example	Destination
CDM-ID	6f44745f-dbc2-4dd6-8913-06d34568035d	Mandatory, will be matched in CDM
WFO-ID	wfo-0001302541	Mandatory, will be added to CDM
IPNI-ID	77142811-1	Optional, will be added to CDM
Name	Nepenthes amabilis Wistuba , Gronem. , Micheler , Marwinski , Gieray , Coritico & V.B.Amoroso	Optional (control only)

2.3 Import of names without classification – currently accepted items in a DwC-A import

2.4 Additional items for import of names with classification

3 Pre-processing of WFO data

Preprocessing can be done with various tools (e.g. Open Refine or Access). The aims are

- To improve the data quality before import
- To identify the publication type of the name (article in journal or book)
- To match the WFO data to another dataset (e.g. Tropicos, see section 3.1)

3.1 Using the Tropicos bulk matching tool

3.1.1 Tropicos output from bulk name matching (examples)

Tropicos allows to upload a text file to http://www.tropicos.org/NameMatching.aspx using the following specification:

The file should include tab delimited columns with the first row containing headers. Column headers should match the column names FullnameWithAuthors, FullnameNoAuthors, and SourceID. At least one of the name columns is required. The order of these columns does not matter.

The output returned has the following format (here transposed and with 2 examples):

FullnameWithAuthors	Ochlopoa annua (L.) H. Scholz	Poa annua L.	Input (the author abbreviations should be adapted to the Tropicos „standard“ – no space between initials, but space between initials and (abbreviated) last name.
FullnameNoAuthors	Ochlopoa annua	Poa annua	Input
SourceID	123	124	Input – use the WFO-ID, where possible
OutputNameID	50231428	25509881	Tropicos ID
OutputHowMatched	FullnameWithAuthors	FullnameWithAuthors	Matching statement (to be ignored in import to CDM)

OutputFullNameWithAuthors	Ochlopoa annua (L.) H. Scholz	Poa annua L.	Matching name in Tropicos (incluiding authors)
OutputAbbreviatedTitle	Ber. Inst. Lanschafts-Pflanzenokologie Univ. Hohenheim Beih.	Sp. Pl.	Nomenclatural title abbreviation (of book or journal)
OutputCollation	16: 58	1: 68	Volume and details concatenated (to be ignored in import to CDM)
OutputVolume	16	1	Volume w/o Issue
OutputIssue			Issue
OutputPage	58	68	Page
OutputTitlePageYear	2003	1753	This is either the actual year for the publication, or, if there is an entry in the next field, it's the to-be-corrected publication year
OutputYearPublished			This is empty, except where the OutputTitlePageYear needs to be corrected.
OutputNomenclatureStatus	No opinion	illegitimate	Nomencl. status – ignore if empty, „No opinion" or „legitimate"
OutputBHLLink		http://www.biodiversitylibrary.org/openurl?pid=title:669&volume=1&issue=&spage=68&date=1753	URI link to BHL
OutputBatchID	8191	8191	Can be ignored.

Another example:

FullnameWithAuthors	Limonium capense L. Bolus	Limonium drepanostachyum Ikonn.-Gal.	
FullnameNoAuthors	Limonium capense	Limonium drepanostachyum	
SourceID	123	124	
OutputNameID	25400115	25400342	
OutputHowMatched	FullnameWithAuthors	FullnameWithAuthors	
OutputFullNameWithAuthors	Limonium capense L. Bolus	Limonium drepanostachyum Ikonn.-Gal.	
OutputAbbreviatedTitle	S. African Gard.	Trudy Bot. Inst. Akad. Nauk S.S.S.R., Ser. 1, Fl. Sist. Vyssh. Rast.	
OutputCollation	1934, xxiv. 124, in obs., 129, in adnot.	2: 267, f. 4	May have to be parsed into following fields, in the first example:
OutputVolume		2	1934(24)
OutputIssue			
OutputPage	1934	267	124 in obs., 129 in adnot.
OutputTitlePageYear		1936	
OutputYearPublished			1934
OutputNomenclatureStatus	No opinion	No opinion	
OutputBHLLink			
OutputBatchID	8192	8192	

3.1.2 Format specification and data cleaning for Platform Import of Tropicos data

As pointed out in the example data (see above), the table containing the output from Tropicos Bulk Matching needs to be controlled and partially transformed by:

- uniting OutputVolume and OutputIssue („34“, „2“ => „34(2)“).
- where the reference data exist only in the OutputCollation, distribute to Volume/Issue and Page (=Detail)
- adapt OutputTitlePageYear / OutputYearPublished – the former is used for the „normal“ „true“ publication year, but if there is an entry in the latter, it was corrected. Distribute correctly to the respective fields.
- Add a collumn to indicate, if the reference title refers to a book („B“) or to a journal (article, „A“)[4].
- Optional: Link with WFO-Download to include IPNI IDs and reference title where not present in Tropicos output.
- Optional but preferred: Link with WFO-Download to include authors abbreviated according to IPNI standards.
-

[N.B.: Character encoding for Platform import: UTF-8]

Specification of the table for Platform import of names:

Column-name	Explanation
WFO-ID	The WFO name identifier (Source-ID in the Tropicos input and output)
IPNI-ID	The IPNI name identifier (optional, taken from WFO provided file)
Tropicos-ID	The OutputNameID in the Tropicos output
WFO-Name	The original full name from WFO output (or the FullNameWithAuthors fed to Tropicos)
NomRefTitle	OutputAbbreviatedTitle in Tropicos output (standard abbreviation of book or journal)
NomRefVolume	Volume and Issue (if applicable), Format: Volume(Issue)
NomRefDetail	Collation of page(s), page range, f. or t. designations
NomRefYear	The actual year of publication
NomRefStatedYear	The year on the title page, if different from actual publication, cite as „as yyyy“
NomStatus	Nomenclatural status
Protologue-URI	URI linking to the original publication in BHL
NomPubType"	A (Article – i.e. NomRefTitle designates a periodical) or B (Book)

[4] WGB has a VBA function for identifying Articles, based on a collection of text snippets (which do need to be improved with new datasets).

4 WFO Backbone output DwC-A Metadata File

```
<?xml version="1.0"?>
<archive xmlns="http://rs.tdwg.org/dwc/text/">
        <core encoding="UTF-8" linesTerminatedBy="\n" fieldsTerminatedBy="\t" fieldsEnclosedBy="&amp;quot;"
ignoreHeaderLines="0" rowType="http://rs.tdwg.org/dwc/terms/Taxon">
                <files>
                        <location>classification.txt</location>
                </files>
                <id index="0"/>
                <field index="0" term="http://rs.tdwg.org/dwc/terms/taxonID"/>
                <field index="1" term="http://rs.tdwg.org/dwc/terms/scientificNameID"/>
                <field index="2" term="http://rs.tdwg.org/dwc/terms/scientificName"/>
                <field index="3" term="http://rs.tdwg.org/dwc/terms/taxonRank"/>
                <field index="4" term="http://rs.tdwg.org/dwc/terms/parentNameUsageID"/>
                <field index="5" term="http://rs.tdwg.org/dwc/terms/scientificNameAuthorship"/>
                <field index="6" term="http://rs.tdwg.org/dwc/terms/family"/>
                <field index="7" term="http://rs.tdwg.org/dwc/terms/subfamily"/>
                <field index="8" term="http://rs.tdwg.org/dwc/terms/tribe"/>
                <field index="9" term="http://rs.tdwg.org/dwc/terms/subtribe"/>
                <field index="10" term="http://rs.tdwg.org/dwc/terms/genus"/>
                <field index="11" term="http://rs.tdwg.org/dwc/terms/subgenus"/>
                <field index="12" term="http://rs.tdwg.org/dwc/terms/specificEpithet"/>
                <field index="13" term="http://rs.tdwg.org/dwc/terms/infraspecificEpithet"/>
                <field index="14" term="http://rs.tdwg.org/dwc/terms/verbatimTaxonRank"/>
                <field index="15" term="http://rs.tdwg.org/dwc/terms/nomenclaturalStatus"/>
                <field index="16" term="http://rs.tdwg.org/dwc/terms/namePublishedIn"/>
```

In the data, we have included 3 unmapped columns with what we have from the info you requested: collation, pages and date.

```
                <field index="20" term="http://rs.tdwg.org/dwc/terms/namePublishedInID"/>
                <field index="21" term="http://rs.tdwg.org/dwc/terms/taxonomicStatus"/>
                <field index="22" term="http://rs.tdwg.org/dwc/terms/acceptedNameUsageID"/>
                <field index="23" term="http://rs.tdwg.org/dwc/terms/originalNameUsageID"/>
                <field index="24" term="http://rs.tdwg.org/dwc/terms/taxonRemarks"/>
                <field index="25" term="http://purl.org/dc/terms/created"/>
                <field index="26" term="http://purl.org/dc/terms/modified"/>
                <field index="27" term="http://purl.org/dc/terms/references"
                     default="http://www.worldfloraonline.org" />
                <field term="http://rs.tdwg.org/dwc/terms/kingdom" default="Plantae"/>
                <field term="http://rs.tdwg.org/dwc/terms/nomenclaturalCode" default="ICBN" />
                <field term="http://purl.org/dc/terms/bibliographicCitation"
                     default="CC0 1.0 Universal (CC0 1.0). https://creativecommons.org/publicdomain/zero/1.0"/>
                <field term="http://purl.org/dc/terms/rightsHolder"  default="World Flora Online Consortium"/>
                <field term="http://purl.org/dc/terms/license"
                     default="https://creativecommons.org/publicdomain/zero/1.0"/>
                <field term="http://purl.org/dc/terms/rights" default="CC0 1.0 Universal (CC0 1.0)." />
        </core>
        </archive>
```

5 EDIT Platform Terminology

[Text from the general Platform Manual]

The ***Classification*** is the uppermost hierarchical element in the Platform's handling of taxa. Several classifications can reside in a single database. This is useful, for example, when there are alternative views on taxon circumscriptions (e.g. in the treatment of the genera Hieracium and Pilosella in the Cichorieae, see http://cichorieae.e-taxonomy.net/).
An accepted (correct) name nested within a classification designates a **Taxon Node,** representing a taxon in a given classification. If a taxon node is assigned to a taxon of a higher rank, the latter is referred to as the **parent taxon**, the former as the latter's **child taxon**.
A **Taxon** is a taxonomic group with the data that define its circumscription and describe its properties. The circumscription of the taxon is indicated by means of a circumscription or concept reference **("sec.-" or "secundum-" reference)**, normally a bibliographic reference clarifying the distinction of this taxon from other taxa. One and the same taxon may occur in several classifications, but it is also possible that two different taxa (taxon concepts) carry the same name in separate classifications. In themselvers, classifications should be taxonomically consistent, i.e. every name should only occur once (as a taxon name or a synonym) in a given classification (except when cited as a misapplied name).
Scientific names (as well as the names given to pseudotaxa) are assigned to records representing taxa, syonyms or misapplied names. **Ranks of names** follow the hierarchy defined in the nomenclatural codes. However, you are free to insert further ranks at any place in the tree thus forming new or mixed hierarchical levels.

6 Import from other sources

6.1 Procedure X: Tropicos Bulk Data Matching

- Preparing a table for Tropicos Bulk Data Matching
    - Extract the 3 data fields for Tropicos Bulk Data Matching[5] from the table obtained from the WFO Data Center
    - Transforming the author names from IPNI (WFO-) Standard to Tropicos "standard"[6]
    - Optional: searching for species and infraspecific names that belong to the group using IPNI or other sources, adding them, where missing to that table (e.g. recently published names)
    - Optional at this step, but must-do before import into the Platform: Exclude those names already present in the Platform database that received a WFO-ID.
    - [Note: character encoding: "Western European (Windows)]
- Bulk upload to Tropicos; revise non-matching results to (partly) identify misspellings etc.
- Bulk upload to Tropicos; receive results table with additional data for nomenclatural reference etc.
- Data cleaning of the result set nomenclatural references (see critical points in the tables below)
- Identify the taxon node in the Platform treatment, to which the data should be added. This can be a pseudonode, e.g. "Additional-WFO-Data" under the "Plumbaginaceae" node.
- Importing of names and other data into the Platform (initially all names will become Taxa in one level below the taxonomic group names taxon node or the respective pseudotaxon
- Treatment of the data in the Platform (assign to correct higher taxon, change to synonym, etc.)

[5] SourceID (=wfo-ID), FullNameNoAuthors, and FullNameWithAuthors

[6] WGB has a VBA function that achieves that.