Notes on the CITES reference document on Nomenclature for Appendix II listed orchids

General explanation

A comparison was conducted between the Nomenclatural Reference Document adopted at CoP 19 and a dataset from 27.10.2023 from the database Plants of the World Online. The outcomes of the comparison are compiled and presented in an Excel spreadsheet ("main_comparison_orchid_AppII_reference_DE.xlsx").

All data in columns B – F are derived from the Appendix II orchid reference pdf file from section "Part II: Binomials in current usage" from pages 509 (514) to 887 (892) and the column´s title is marked with an appended "_reference". These pages were chosen, because the names could be automatically extracted more easily than from the other parts of the document, but we can not guarantee that the extraction process worked flawlessly. Species names were extracted using pdfminer.six [1], subsequent comparisons were performed using pandas [2] and general programming logic using python [3].
To compare the data in the orchid reference we used data obtained from Plants of the World Online (PoWO), that should be in principle, the same data as wcsp that was used to create the reference, although at a different point of time. Data from PoWO is found in columns H to Q, where column titles could be confused with the ones from the reference, "_PoWO" was appended to the title. The complete dataset from PoWO of all orchids was retrieved on 27.10.2023 including "taxonomicStatus" and "namePublishedInYear" ("year_PoWO") and was reduced to all orchid names that has either an accepted name in one of the genera that is covered by the reference or a synonym to a name of those genera or is a synonym with a genus name of those genera but is considered synonym to a species of a genus that is not included in the reference.

Columns G and R were hidden as they are only needed for the comparison to bridge between different formats of the datasets.

Homonyms

Unfortunately, homonyms could not be properly included in the analysis, they do not match their counterparts from the dataset acquired from PoWO. They are included in the final dataset but should be excluded from all subsequent analyses by setting "reference_homonym" and "powo_homonym" to FALSE. At this time, we cannot recommend how to handle homonyms. We suggest, however, to mark homonyms in future

reference documents to emphasize the difficulties homonyms might cause in implementation.
To include homonyms in the full analysis additional programming logic would be needed to be implemented but was not done due to time constraints.

Year of publication
In two cases, where the year of description is not clear there is a character missing and the two years have been fused:

Angraecum maheense Schltr. ex Diels (18981899 publ. 1922)
Bulbophyllum cootesii M.A.Clem. (19992000 publ. 1999)

These should be corrected for clarity.

For several names the year of publication is not the same as the one published by Kew and might be erroneous. These cases can be reviewed by setting “year_match” to FALSE and filtering out empty cells from “name_reference” and “name_PoWO”.
In a few cases a year is not given in the reference document (at least not in part II) and is marked with “NaN”. Only in four cases where the year of publication is missing there is one given in PoWO (filter for year_reference is NaN AND year_PoWO is not empty AND not NaN).

Special characters
Several special characters are embedded in some species names and or author names that are neither searchable nor easily machine-readable. The following is a dictionary in json-format that represents a translation table from the special characters that were read from the pdf to the character or sequence of characters they are supposed to represent:
{'fl': 'fl', 'fi': 'fi', '´e': 'é', '`e': 'è', '´E': 'É', '´o': 'ó', '¨u': 'ü', '´a': 'á', '´i': 'í', '´ı': 'í', '´u': 'ú', '¨o': 'ö', 'ff': 'ff', '^o': 'ô', 'ffi': 'ffi', '¸c': 'ç', '˜n': 'ñ}

The following names are examples that cannot be found because of the special characters which replace “ffi”, “fi” and “ff”:

Aerangis flabellifolia
Aerides affinis
Aerides magnifica

Approximately 1120 names were affected and thus are not searchable in the pdf.

The other characters that cannot intuitively be found are in the author names.

Unplaced names
The reference marks eight names as unplaced, according to PoWO 19 names on the reference are unplaced, some of those names are homonyms. As mentioned under the section “Homonyms”, we currently have no suggestion how to handle them and thus were excluded from the analysis.

The following name is unplaced according to the reference, but according to PoWO can be placed:

Vanda flavobrunnea Rchb.f.

313 additional unplaced names were not included in the reference. There is no need to include them, but we were wondering why some unplaced names were included in the reference but most were not.

Unplaced names from PoWO and all homonyms are excluded for the following analyses:

Author mismatch
By filtering “author_match” == FALSE and filtering for non-empty cells in “name_PoWO” AND “name_reference” one can find the names where the author names do not match exactly, which are 252 entries.

Year mismatch
By filtering “year_match” == FALSE and filtering for non-empty cells in “year_PoWO” AND “year_reference” one can find the names where the year of publication do not match exactly, which are 358 entries. Year entries where the format was “1972 publ. 1973” where not analysed in more detail.

Names status without consensus
Filtering for “status_consensus” == FALSE will display all names that are accepted by one source and synonym by the other source as well as names that are present in one dataset but not in the other.
To only find entries where the names match but there is no consensus on the status (accepted name or synonym) one needs to filter for “status_consensus” == FALSE AND “name_reference” is not empty AND “name_PoWO” is not empty, the result are 209 entries.
These might be errors in the reference, but the status of these cases might also have changed since the creation of the reference. Nevertheless, these

cases might cause confusion over the correct status when working with the reference.

Names only present in one dataset

To find names that are only present in the PoWO dataset but not in the reference, one might need to filter for entries where “status_consensus” == FALSE to decrease the overall dataset and be able to apply subsequent filters. On my machine with my Excel version I could not apply the following filters without reducing the dataset first. To find names that are only present in the reference, one needs to deselect empty cells from “name_reference” and deselect all fields from “name_PoWO” and select the empty cells.
170 names are present in the reference that were not found in PoWO.
At least 118 cases of those did result from different spellings between both datasets which can be reviewed in the additional file ‘misspelled_names.xlsx’. This file contains suggestions which name from PoWO is the closest to the name in the reference, that had no exact match in the main comparison. The most similar name in the PoWO dataset was found using RapidFuzz [4], utilizing the RapidFuzz process.extract() module. The similarity_score provided was calculated by multiplication of the RapidFuzz ratio with the Levenshtein distance. The more similar the names are, the lower the similarity_score.
To filter for names that are present in PoWO but not in the reference the filter for empty cells in “name_PoWO” needs to be deselected and in “name_reference” only the empty cells must be selected.
The result is 3655 names, most of them are synonyms.
There are 898 accepted names within this subset, whereas most of them (759) have a publication date later than 2017 and thus might have been published after creation of the dataset for this reference. That still leaves 139 names that are missing from the reference.
With the same filter applied, we identified 2594 synonyms missing from the reference and 170 published after 2017 and 8 with no year of publication provided.

General notes on format

As a general note on ease of use for the orchid Appendix II reference would be to adopt a simple format like json and maybe also csv, which might be accessible for more people. Large datasets should be provided in a machine-readable format to allow easier comparison with widely accepted databases (including speciesplus.net) and also management and maintenance of Member States´ national databases.

References:

[1] Shinyama, Y., *et al*. (2023). pdfminer.six (Version 20231228) [Software]. Available from https://github.com/pdfminer/pdfminer.six

[2] McKinney, W. *et al*. (2023). pandas (Version 2.1.4) [Software]. Available from https://pandas.pydata.org/

[3] Python Software Foundation. (2023). Python Language Reference, version 3.10.13 Available from https://www.python.org

[4] Bachmann, Max (2023). RapidFuzz (Version 3.6.1) [Software] Available from https://github.com/rapidfuzz/RapidFuzz