# Release notes for Flora Malesiana, Flore du Gabon, and Flora of the Guianas XML files

ver. 1.0

Thomas Hamann

**Copyright:** Document copyright © Thomas Hamann/Naturalis Biodiversity Center 2016. This document is licensed under a Creative Commons Attribution-ShareAlike 3.0 Unported (CC BY-SA 3.0) license.

## Table of Contents

# Introduction

This document provides an overview of XML files that exist for Flora Malesiana, Flore du Gabon, and Flora of the Guianas, together with a summary of the known issues.

As the Perl scripts were developed at the same time as the volumes were marked up, there are differences in the quality of mark-up and degree of atomisation that are concurrent with the development state of the scripts at that time. The implication of this is that volumes that were marked up first have less atomisation and may be of qualitatively lower standards than the most recent volumes. This is a mere consequence of the chosen approach and is very much related to the time-consuming process of obtaining a better understanding of the contents of these floras. Preventing it would have required a complete understanding of all of the problems and other quirks present in these works and their interaction with the XML mark-up and the database import, prior to actually starting the development of the scripts and adding mark-up to the flora volumes, which for all practical purposes is unfeasible.

The volumes that were marked up prior to 2016 were partially upgraded to bring them closer to current standards once the mark-up of Flora Malesiana was completed.

## *General notes*

- Some keys referred to taxa not present in a volume. These taxa were added to the volume.
- Some infraspecific taxa, and sometimes also other taxa, were only mentioned in the text of another taxon. These were separated from their parent taxon whenever this was needed, i.e. the taxon name was bolded or otherwise indicated to be of enough importance to warrant separation.
- Certain specific symbols, such as ♂ or ♀, were not recognised properly and were either added in or replaced by their equivalent word. A similar thing applies to fractions.
- OCR errors were corrected whenever spotted.
- A few printing errors were dealt with prior to mark-up, after discussion with resident editors or taxonomic experts.
- There is no generally applicable order for specimen data. Likewise, there is no generally applicable order for the paragraphs in the larger taxon treatments (e.g. family-level).

- Normalisation of data was very sparsely applied, as it is often impossible to verify whether something is an error (compared to an online source[1]) or was *meant* to be that way by the author(s) of the treatment.
- In some cases, text was reordered to be able to better deal with the data contained therein.
- Names located in free text are not marked up, except in Flora of the Guianas.

[1] Which can also be wrong.

# Flore du Gabon

44 XML files, most of single volumes. There are two exceptions: vols. 12 and 17, and vols. 36 and 37, which are two files instead of four because each set treats a single family. Vol. 5bis consists of addenda to vol. 5.

## *Known original print issues*

- Species epithets often start with a capital letter.
- In older volumes, a single list of citations is given for all taxonomic names provided instead of a list of citations per listed synonym. Sometimes each synonym is accompanied by a year indicating which citations belong to which synonym, but this is not always the case. For the mark-up, all citations were placed under the accepted name in these cases.
- In older volumes, it is sometimes unclear whether synonyms are homotypic or heterotypic. For the mark-up, it was assumed that synonyms on separate lines were heterotypic.
- In newer volumes, types are not given.
- Incomplete citation information with missing reference parts sometimes present.
- In vol. 30, figures are referred to which are not actually present. A dummy figure with no contents was added to deal with this.

## *Known mark-up issues*

- The descriptions of vols. 1 to 4 are atomised to a lesser degree than those of the subsequent volumes.
- Vols. 1 to 4 were marked up by hand to a much greater degree than subsequent volumes, as the Flore du Gabon scripts were still in their early development stages at that time.
- Vol. 5bis uses the addendum-specific mark-up. Unlike addenda in other floras it was not merged with its companion volume.

# Flora of the Guianas

16 XML files, all for single volumes of series A. Two files, for vols. 24 and 25, are enhanced duplicates of the standard files for vols. 24 and 25.

For this flora we tried to push atomisation much further than for the other two floras.

## *Known original print issues*

- There are some minor problems with a lack of punctuation in some descriptions.

## *Known mark-up issues*

- Vol. 24 has less advanced description atomisation than subsequent files (but still better than Flore du Gabon and Flora Malesiana).
- The general order (and general ordering method) of specimen data in Flora of the Guianas is different than that in other floras. Changing the order to that of the other floras would have broken the way specimens are indexed in Flora of the Guianas for human users. Therefore it was marked up in its original order.

# Flora Malesiana

32 XML files, most for single volumes (Series I vols. 4-6, 12-21, Series II vols. 1-4) or single parts of volumes (vols. 7p1-2, 7p4, 8p1-3, 10p1-4, 11p1-3). One file for 2/3rds of vol. 9, and one file for two halves of a family coming from two volumes.

Due to the very disjunct manner in which the Flora Malesiana volumes were converted to XML (periods of several months spread over several years interrupted by other periods of several months), this flora suffers most from the issue identified in the introduction. Furthermore, volumes 12 and 13 were marked up entirely by hand prior to the development of the Perl scripts. However, most issues are caused by the manner information is presented in the first place (see big list below).

The few hundred pages of addenda, corrigenda and errata have been merged into the actual volumes prior to mark-up.

## *Known original print issues*

- Citations in older volumes are given in chronological order, with usually no distinction between homotypic and heterotypic synonyms. Types are often lacking. More recently published volumes do make the difference between homo- and heterotypic synonyms, give citations per synonym, and also provide type information.
- There are multiple variant abbreviations for the same publication name (up to a dozen per publication name). Even after the introduction of a standardised

list of publication name abbreviations, the use of multiple abbreviations continued.

  - It is important to understand that the standardised list of publication name abbreviations included in Series I volume 5 is *only* valid for the Flora Malesiana flora; it does *not* consist of the generally accepted abbreviations for publication names as used by botanical libraries.
  - Some authors are *very* inconsistent in their use of abbreviations.

- Some citations and references are problematic because they:
  - refer back to literature previously cited that actually is absent from the volume.
  - refer to dates in the future of the publication date of the Flora Malesiana volume.
  - are missing the year, or sometimes even more.
  - some other reason.
- The inclusion of publication details such as the series, part, issue, etc. numbers is unfortunately very messy, depending on the Flora Malesiana volume (more recently published volumes are better) and author. Multiple formats exist for the same publication.
  - Adding to the confusion are edition numbers that haven't clearly been marked as such (e.g. no "2nd ed." or "ed. 3"). This leads to confusing situations when a publication has both multiple series and editions. The same sometimes happens with sections and appendices.
  - In earlier volumes the part number was sometimes indicated with a superscript number or a number in a different style following the preceding one with no space present. These often did not survive the OCR process and may not have been fixed.
- Nomenclature sometimes includes multiple spelling options for a single taxonomic name.
- Authors seem to have been confused by what is supposed to follow an equal sign (=) in excluded taxa and name types (correct name, basionym, something else…).
  - Additionally, at least one author explicitly marked names as basionyms that cannot possibly be basionyms according to the Botanical Code.

    - Some authors included the equal sign followed by a name in normal nomenclature. We don't know what these are.

- We have not been able to figure out the format used for vernacular names at the end of a taxonomic treatment. The same abbreviations seem to have been used for both the local language and the locality, and the text format (punctuation, order of information) used is extremely inconsistent from author to author and volume to volume.

- Specimen information in older volumes of Flora Malesiana was typically not given in the flora itself, but in separate, stencilled publications with absolutely atrocious printing quality. The only specimen information given in the flora itself is located in figure captions and types. This has been marked up in all files, except for the most recent ones, as the current editor indicated a preference for using online specimen databases because they are more complete.

- There are a few cases of actual printing errors, such as missing information, very obvious typos, and missing pages. These have been fixed whenever possible.

### *Known mark-up issues*

- The problems with references and citations mentioned above has the consequence that some literature references may have been unintentionally incorrectly atomised. This is because the scripts assume that the text format used always follows the format most commonly used in Flora Malesiana.

- Often, text in the flora features references back to earlier literature references or citations using the abbreviation "l.c.". In more recently marked up volumes these have almost always been expanded into the full reference. This is not the case for volumes marked up several years ago.

- Series I vols. 12-20 and Series II vol. 3 still use the deprecated <taxontitle>-element, which is not used in any other volumes. Be aware that the handling of taxon numbers is different when using this element, compared to the later mark-up.

## Other issues

There are a few other issues with the contents of legacy flora data that are not listed in this document because they are of no direct impact on the XML mark-up. Instead, they lead to problems with the interpretation and exchange of the data. These are not discussed in this document, but will be discussed in my future article on the subject.