METHODS AND TECHNIQUES

Detailed mark-up of semi-monographic legacy taxonomic works using FlorML

Thomas D. Hamann,[1] Andreas Müller,[2] Marinus C. Roos,[1] Marc Sosef[1,3] & Erik Smets[1,4]

1 *Naturalis Biodiversity Center, P.O. Box 9517, 2300 RA Leiden, The Netherlands*
2 *Botanic Garden and Botanic Museum Berlin-Dahlem, Freie Universität Berlin, Königin-Luise-Str. 6–8, 14195 Berlin, Germany*
3 *National Botanic Garden of Belgium, Domein van Bouchout, Nieuwelaan 38, 1860 Meise, Belgium*
4 *Plant Conservation and Population Biology, KU Leuven, Kasteelpark Arenberg 31, Box 2437, 3001 Leuven, Belgium*
Author for correspondence: *Thomas D. Hamann, Thomas.Hamann@naturalis.nl* or *tdhamann@ziggo.nl*

DOI http://dx.doi.org/10.12705/632.11

Abstract We present FlorML, an XML schema, specifically designed for the detailed mark-up of highly complicated semi-monographic legacy taxonomic works, such as large Floras and Faunas. We discuss the prerequisites for developing a suitable XML schema, and the limitations presented by the legacy taxonomic works, requirements by stakeholders and the desired output format. Furthermore, we explain how FlorML was deployed to mark up two legacy taxonomic works, *Flora Malesiana* and *Flore du Gabon*, how that deployment was improved by the use of the scripting language Perl to automate major parts of the process, and discuss the issues commonly encountered during the mark-up process. Examples and figures are provided for clarification. Finally, we make suggestions for future research and further developments in the field of biodiversity informatics.

Keywords biodiversity informatics; e-taxonomy; Flora; legacy literature; XML mark-up

■ INTRODUCTION

In recent times increasing access to biodiversity data for further purposes has become a primary focus area for biodiversity research institutes. Ideally, not only should such data be freely available to anyone who needs it, from the citizen scientist wanting to identify an organism to governments wishing to know what species they should protect, but the format should be such that further uncomplicated reuse of the data is possible for online publishing, linking or data exchange, merging and processing of detailed data for the purpose of data-mining. Such a release of data will hopefully increase the versatility of taxonomic work and therefore revitalize the field (Marhold & al., 2013). Consequently, various initiatives for the development and population of databases including information on the many different types of resources hosted by biodiversity research institutes are underway and gaining more widespread acceptance (e.g., extraction of morphological leaf characters, Corney & al., 2012; herbarium specimen digitization, Barber & al., 2013).

However, the main source of biodiversity data, specific to a particular area or collector, is in publications, distributed in print, known as “Legacy taxonomic works”. These works cover the better part of four centuries and are referenced by various branches of science, commercial companies, educational institutes, various non-governmental organisations and government entities, not all in the area of biology (Thessen & Patterson, 2011). Individual citizens may also be interested in biodiversity data, for example to successfully identify a pest in their garden or discover what exotic plants they photographed during their holidays.

In the last 15 years many of these works have been digitised and made available through online archives, such as the Biodiversity Heritage Library, which has digitized almost 43 million pages of biodiversity literature as yet (March 2014). Increasingly many journals are publishing on-line as well; for example, *ZooKeys* (Pensoft Publishers) publishes species descriptions online with advanced markup of descriptions and links to EOL, SpeciesID, etc. Unfortunately, in most cases immutable file formats, such as PDFs, are used, which are only easily accessible to human readers, not computers. Much of the data is contained in publications that have not yet been digitised and that may be quite difficult to access. Data is often published in large, heavy, or otherwise unwieldy books, or highly scattered in taxonomic articles. Access may be limited due to expensive institutional journal subscription fees, the work being out of print, in a dire physical condition, or having unfamiliar or foreign languages. To disclose all this information in a satisfactory manner, the text of legacy taxonomic works needs not only to be digitised, but also converted to a suitable, machine-readable format. Only then, when the taxonomic data has been made available on the web, can data from external resources be linked to complete missing information both for individual taxa and for the work as a whole. Prior to that, several problems with the legacy taxonomic data have to be resolved, as they generally are very variable and not normalised nor standardised. Furthermore, making taxonomic contents available on the internet suffers from various

Received: 30 May 2013 | returned for revision: 10 Oct 2013 | revision received: 27 Nov 2013 | accepted: 9 Dec 2013 | published online ahead of inclusion in print and online issues: 17 Mar 2014 ||

bottlenecks both in workflows and in the actual linking processes (Feigenbaum & al., 2007; Hagedorn, 2007; Cui, 2008b; Penev & al., 2011; Thessen & Patterson, 2011). This second step and its associated problems will not be extensively discussed in this article, although some of the issues that contribute to the problems of linking data will be discussed. Currently, an EU-funded project (pro-iBiosphere (http://www.pro-ibiosphere.eu/) is underway, coordinated by Naturalis Biodiversity Center, which aims at delivering recommendations and strategies to resolve the aforementioned problems.

To digitise a printed document, the currently most commonly used method involves scanning it, followed by Optical Character Recognition (OCR), ideally resulting in a digital version of the taxonomic treatment text including all of its formatting. However, deteriorated paper, non-standard fonts and special symbols, print quality issues, document age, and low resolution scans can all negatively influence OCR quality (Freeland, 2011).

Once digitised, the text of a taxonomic work can be augmented with contextual information and made machine-readable. The technology used for this is called XML, eXtensible Mark-up Language (Quin, 2010b). The additional information for a given piece of text can be precisely defined using XML elements (small pieces of text between angle brackets). XML not only gives information on the document structure, but can also be used to further split up (“atomise”) text to provide access to low-level data. For example, a taxonomic description is marked up with a specific set of XML elements, while the various characters it contains are each marked up with their own set of XML elements that precisely indicate what each character is. Such XML mark-up can be used for data-mining purposes, but also to construct a multi-access key for a specific taxonomic group. XML elements can have attributes to define specific properties, e.g., to indicate the presence or absence of a certain character. Documents enriched with XML elements are also suitable for preparing legacy taxonomic data for import into a database system. To ensure that the XML is consistently applied to any taxonomic work an XML schema should be used which defines the constraints for each XML element and attribute, e.g., the data type of an XML element. It also defines how each XML element relates to other elements at the same level, at a higher level (“parent elements”) or at a lower level (“child elements”). There are several kinds of XML schemas; in our case we used the XSD (XML Schema Definition) format (Biron & Malhotra, 2004; Fallside & Walmsley, 2004; Thompson & al., 2004, Quin, 2010a).

Several XML schemas are currently in use for the mark-up of legacy taxonomic publications. Of those schemas meant for the mark-up of taxonomic treatments in (legacy) journal articles and books, TaxonX and TaXMLit are the best-known. TaxonX is suitable for high-level mark-up of the structure of taxonomic treatments and the phrase-level mark-up of specific types of contents such as taxonomic names and localities. It also has the ability to link to external resources and mechanisms for semantic normalization (Penev & al., 2011). TaXMLit supports all types of contents commonly found in taxonomic literature and the treatments within, except for characters in the descriptions. Atomisation of many types of contents is possible (Weitzman & Lyal, 2004). However, both have their limitations. TaxonX’s abilities to support and atomise various types of taxonomic data are limited, while TaXMLit often expects that structural elements are placed in a fixed order. Furthermore, in many projects XML mark-up is limited to the document structure, with little to no atomisation (Penev & al., 2011).

For our purpose, mark-up of two large multi-volume Floras published during some 50 years, we needed an XML schema that had the ability to deal with large and complex legacy taxonomic works with very variable document structure and contents, combined with stakeholders requesting far-reaching atomisation. In this article we present an XML schema capable of dealing with such works and requirements: FlorML (Flora Mark-up Language). We discuss problems encountered during development and deployment of the XML schema and possible solutions. Furthermore, we discuss how to automate a large part of the mark-up process, as ideally XML mark-up of a document is a fully automated process. Finally, we provide suggestions for further improvement of digitisation and mark-up of printed taxonomic information.

■ MATERIALS AND METHODS

The development of FlorML started with a six-month pilot project in May 2010, eventually aiming at entirely digitising the semi-monographic *Flora Malesiana* (Roos & al., 2011). This pilot project intended the mark-up of *Flora Malesiana* in order to use it in the EDIT (European Distributed Institute of Taxonomy; http://www.e-taxonomy.eu/) CDM (Common Data Model) database system. This database for all types of data that taxonomists usually produce was developed by the Botanical Garden and Botanical Museum Berlin-Dahlem (BGBM) and its EDIT partners in London, Paris and Brussels. It is part of the EDIT Platform for Cybertaxonomy (http://cybertaxonomy.eu/; Berendsohn, 2010; Venin & al., 2010; Berendsohn & al., 2011) and facilitates online publishing, joining data with external resources, and key creation, amongst others. The pilot project was subsequently extended to include *Flore du Gabon*, which allowed us to test and deploy FlorML on another Flora.

Legacy taxonomic works used. — *Flora Malesiana* is a systematic account of the flora of Malesia, the plant-geographical region spanning six countries in Southeast Asia: Indonesia, East Timor, Malaysia, Singapore, Brunei Darussalam, the Philippines, and Papua New Guinea (Van Steenis, 1947). It consists of two series; Series I for angiosperms and Series II for ferns, with a total of 23 published volumes. Furthermore, nine CD-ROMs focusing on Leguminosae and Orchidaceae were published together with ETI Bioinformatics, and an interactive key in collaboration with the Royal Botanic Gardens Kew. The *Flora Malesiana* project started in 1948 and as of 2012 covers about 25% of the currently estimated 40,000 plant species of Malesia. *Flora Malesiana* Series I volumes 1–14 and Series II volumes 1 and 2 had to be scanned and OCR applied to them, while all subsequent volumes were digitally available in Adobe PageMaker and InDesign format.

Flora Malesiana has highly detailed taxon descriptions including various features such as morphology, anatomy, palynology, biochemistry, amongst others. It often features very detailed synonymy. The first few volumes contain biographies and general chapters on ecology, biogeography, and the practices of collecting in tropical countries. Some volumes contain errata for taxa described in previous volumes.

In nearly 60 years, *Flora Malesiana*'s format has regularly changed, affecting both the general appearance of the *Flora* (e.g., fonts, number of columns) and text formats used in the *Flora* (e.g., literature citations). Despite this, the various volumes are generally highly and consistently structured, especially since 1997 due to having been produced through digital means with stricter editorial control.

Flore du Gabon, produced in French, counts 45 volumes covering about 2650 species of the ca. 5200 species of Gabon (Sosef & al., 2006). The first volume was published in 1961. Volumes 1–37 had to be scanned and OCR applied to them, while all subsequent volumes are available as Adobe InDesign files.

Flore du Gabon has succinct nomenclature, descriptions, and other features. It only occasionally contains non-taxonomical information. In earlier volumes, especially those published prior to 1965, *Flore du Gabon* adopts a narrative writing style for the text, mixing taxonomic descriptions with other features such as distribution and ecological information. Usually, type information does not immediately follow the nomenclature, but is placed after or even inside the taxon description. This means the document structure is very loose in general. From 2008 onwards, a new format is used that is much more structured and does not name types.

In both cases, we have aimed at making the legacy taxonomic treatments contained in the described Floras fully available through an EDIT CDM Data portal.

Mark-up work prior to XML schema development. — FlorML's roots can be found in the XML mark-up procedures as developed by Kirkup & al. (2005) for the Royal Botanic Gardens, Kew exemplar e-Flora, *Flora Zambesiaca* (see http://apps.kew.org/efloras/). This XML schema supports basic mark-up of most of the contents found in a taxonomic treatment, such as dichotomous keys and a limited number of features: description, distribution, habitat and ecology, chromosomes, uses and vernacular names. Only nomenclature and part of the description characters were atomised, while literature references were not. All remaining text was considered to be notes.

Although this schema is perfectly suitable for taxonomic works with a simple structure, several problems were found when marking up *Flora Malesiana* with it. Most were related to the more complicated structure and greater abundance of topics discussed in that Flora, others had to do with more advanced stakeholder requirements.

XML schema development and description. — The text structure and types of content of the printed works involved have to be analysed prior to developing an XML schema suitable for marking up legacy taxonomic works. The chosen approach is a semi-structured one where not only the basic document structure is marked up, but various parts are atomised in further detail to improve their usefulness for future applications.

A complicated Flora often has specificities that are not noticed when the mark-up only covers the document structure at a high level, with no to little atomisation, but do become important at more granular levels, such as full nomenclatural and reference atomisation. Then they may well interfere with proper mark-up. Some, related to taxonomic nomenclature, are explained in the various nomenclatural codes, but others are unofficial taxonomic indications or content specific to an author, such as taxonomic annotations. Most of the former can be accommodated up to a certain degree, but author-specific content can be a serious problem, especially if the author is inconsistent within a single taxonomic work. To gain a better understanding of these sometimes mystifying and idiosyncratic issues, experienced taxonomists were consulted. Furthermore, bold or italic text formatting not always has the same "meaning", while references to figures, tables, or footnotes may be found anywhere, including in keys. For the FlorML XML schema design this meant that there should be child elements present that can take care of such contents, whenever and wherever present.

To further complicate matters, the contents of legacy taxonomic works are not limited to taxonomic treatments. There are two more types of contents present in legacy taxonomic works: non-taxonomic text such as general introductions to a biogeographical area or biographies of taxonomists, and errata to prior volumes. Furthermore, each taxonomic work also contains metadata, such as ISBN, publication type, publishers, abstracts, etc.

Besides the structure of the legacy taxonomic work, two other factors were very important for the FlorML XML schema design.

First, various types of users have different needs regarding taxonomic data. Field users will prefer descriptive characteristics that allow them to identify certain species; taxonomists working on a new treatment of a certain taxonomic group will also appreciate extensively atomised synonymy; ecologists want descriptive habitat, ecological, and distribution data; finally, ethnobotanists will appreciate data on the use of taxa by local populations and vernacular names, as well as anything related. Furthermore, institutional users and publishers of online Floras may desire to easily construct a more complete Flora by having other databases supply the data missing from taxonomic treatments, especially the descriptive parts of those treatments. This may enable further novel data analyses that could reveal previously unknown patterns within biodiversity data or the construction of advanced interactive multi-access keys. To make this possible, the schema should support atomisation of large chunks of data into much smaller parts.

Second, the feedback from importing each *Flora Malesiana* volume into the EDIT CDM database was used to make decisions regarding changes or additional features. The EDIT CDM database system's design imposed certain constraints on the design of the FlorML XML schema. In practice, this feedback consisted of many exchanges between staff at the Botanic Garden and Botanical Museum Berlin-Dahlem (the EDIT CDM host) and Naturalis Biodiversity Center discussing both potential and real issues with the XML schema and its implementation. For example, taxonomic names should be atomised as far as possible, but the EDIT CDM database only supported atomised names in the current binomial format, meaning that names that do not adhere to that format should be marked up separately.

The initial approach was to combine all of the above requirements into a single XML schema based on that of Kirkup & al. (2005) that would be sufficiently flexible to mark up anything in *Flora Malesiana*, using Altova XMLSpy (http://www.altova.com/xmlspy.html). This, unfortunately, led to a very complex XML schema with many elements duplicating functionality. This initial schema was then refactored into a more practical version by carefully examining which elements could be combined, and using XML attributes to differentiate between different types of contents. For example, the initial schema had separate elements for every type of descriptive feature of a taxon, leading to a total of 21 elements, each with child elements that had almost exactly the same functionality. These different descriptive feature elements were collapsed into a single element <feature> with an attribute describing the type of feature (Fig. 1). In the initial schema, the metadata were marked up using proprietary XML elements. This was replaced by an already existing XML schema for publication metadata, MODS (Metadata Object Description Schema: http://www.loc.gov/standards/mods/) in FlorML.

Further development of the FlorML schema consisted of ensuring that functionality was present where needed, using a data-driven approach. Required additional taxonomic features and characters were added upon encounter. When the first volumes of *Flore du Gabon* were marked up, multi-language support was added to FlorML to differentiate between descriptions in *Flore du Gabon* that were both in French and Latin. The initial implementation for basic character atomisation proved insufficient to properly mark up some descriptions of species with separate male and female flowers. These descriptions used the same characters for each of the two genders (Fig. 2A), leading to issues during the import into the CDM database system because characters were seemingly duplicated despite actually being different (Fig. 2B). This was resolved by introducing sub-characters (Fig. 2C).

Based on a small-scale survey among Naturalis Biodiversity Center staff, we also added some optional basic functionality for further atomisation beyond the level of whole (sub-) characters, e.g., mark-up of the diameter of a flower. This is currently not deployed yet, pending solutions for issues that limit or prevent further atomisation, such as OCR errors and the very variable nature of language. Figure 3 shows the character mark-up abilities of FlorML.

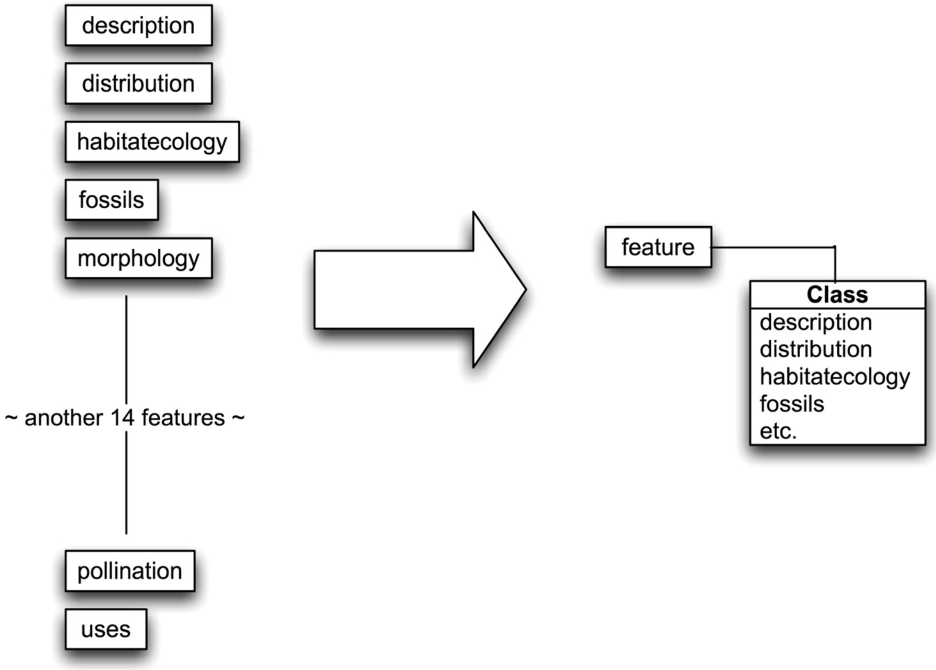
Currently, FlorML subdivides a taxonomic publication into metadata, a treatment part, non-taxonomic text, and errata. The treatment part consists of one or more taxon treatments. Each taxon is subdivided into nomenclature, keys, descriptions, literature, etc., some of which are fully atomised. Figure 4 shows a simplified representation of how FlorML subdivides a document, listing only the more common content types.

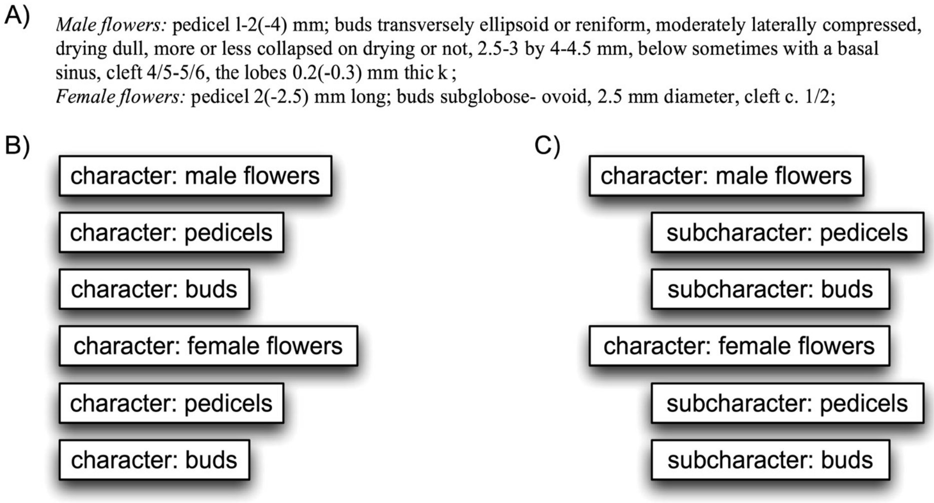
Page numbers, tables of contents and indexes are not used for document structuring in the FlorML XML schema. They can be instantly automatically generated for a printed version of a databased electronic publication.

The mark-up for keys in FlorML supports polytomous keys and certain types of multi-access keys. It is also possible to link from one key (or normal text) to another key.

FlorML supports mark-up of homotypic and heterotypic synonyms, homonyms, basionyms and a variety of types and specimens. Each of these can be fully atomised according to

Fig. 1. Collapsing all possible feature elements into a single element with an attribute called "class" listing the various possible features.




A)

Male flowers: pedicel l-2(-4) mm; buds transversely ellipsoid or reniform, moderately laterally compressed, drying dull, more or less collapsed on drying or not, 2.5-3 by 4-4.5 mm, below sometimes with a basal sinus, cleft 4/5-5/6, the lobes 0.2(-0.3) mm thic k ;
Female flowers: pedicel 2(-2.5) mm long; buds subglobose- ovoid, 2.5 mm diameter, cleft c. 1/2;

Fig. 2. A, fragments of a description of a dioecious species; **B,** when using only one type of character element, characters are seemingly duplicated; **C,** by using both character and subcharacter elements, it is clear which contents belongs to which flower gender.

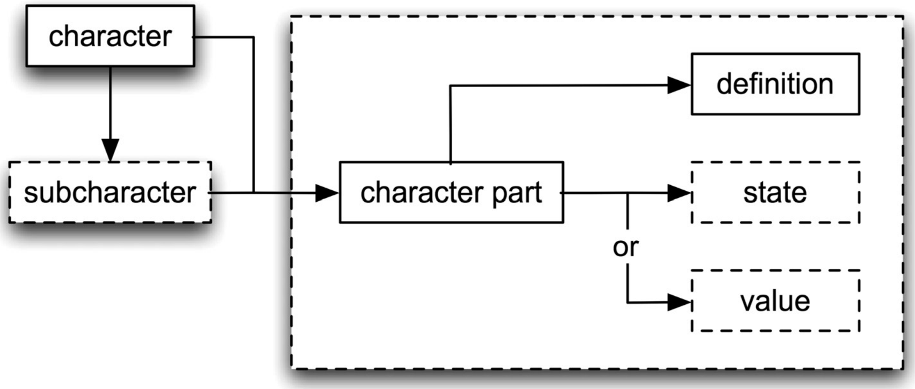


Fig. 3. Character atomisation model. The parts shown within dashed lines are optional.

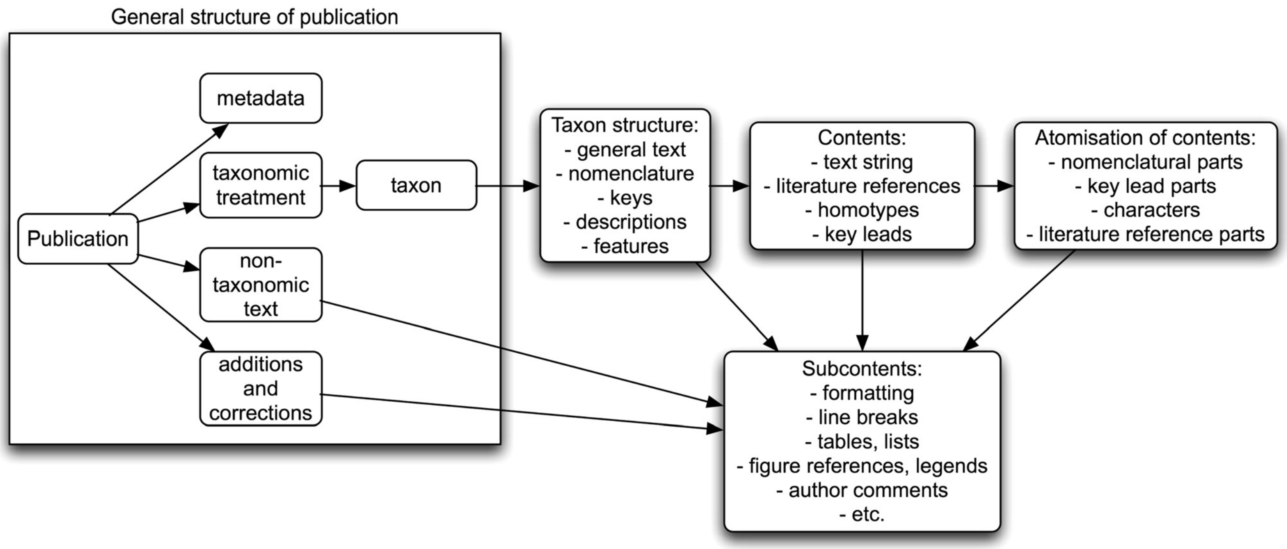


Fig. 4. Simplified schematic representation of FlorML XML schema.

the options given in the most recently available versions of the *ICN* (McNeill & al., 2012) – including fossils (Cleal & Thomas, 2010a, b; Cleal, 2011), *ICZN* (Ride & al., 2000), and *Bacteriological Code* (Lapage & al., 1992; despite some minor terminological differences). This remains to be tested for the latter two codes. Full-text fields for names and types that cannot be atomised are also provided.

Currently, FlorML supports a total of 72 possible taxon features. These include descriptions, geographical distributions, habitat and/or ecology information, anatomy, fossils, chemotaxonomy, phylogenetics and more. The character atomisation model for descriptive characters supports over 1200 possible botanical characters. Although it is currently not deployed due to time constraints, the possibility exists to mark up characters in other features than the taxon description only. All localities mentioned in geographical distributions can be marked up separately, including coordinates and TDWG regions. Vernacular names and languages can also be atomised. Literature references and citations, whether present within the nomenclature or elsewhere, can be fully atomised.

FlorML provides mark-up for lists, tables, figures and footnotes, and elements allowing references to all of the previous to be linked to their targets using unique identifiers. In the case of figures, these identifiers are also repeated in the figure file names. All other common types of contents found in taxonomic treatments, such as headings, treatment writers, text paragraphs, etc. can also be marked up. FlorML also supports the inclusion of author annotations, which are comments that authors of treatments added to a taxonomic treatment for purposes of clarification or to provide additional information.

The information contained in errata is only integrated into the proper taxon prior to XML mark-up when it is explicitly indicated what the change should be. All other cases will require the expert eye of a taxonomist specialized in the taxa concerned, but special mark-up is provided so they can still be included in the XML documents.

FlorML is designed to be very flexible and aims at facilitating mark-up of legacy taxonomic works as much as practically possible. However, there will always be some cases where text will have to be reordered to make correct mark-up possible. For example, when a type is separated from the rest of the nomenclature in *Flore du Gabon*, it will need to be moved to the end of the nomenclature manually.

FlorML is licensed under a Creative Commons Attribution-ShareAlike 3.0 Unported license. It can be found on the GitHub software repository at https://github.com/ncbnaturalis/FlorML. Its design is currently stable, but remains a work-in-progress and as such currently lacks a namespace. Creating one is planned.

In the next part we will explain how FlorML was deployed.

Manual mark-up. — Mark-up using the FlorML XML schema was initially performed using a method similar to that of Kirkup & al. (2005), in which digitised publications are prepared for mark-up by cleaning up the document and adding different styles to each type of contents. Then automated mark-up is performed using the "Find and Replace"-function in Microsoft Word. Unfortunately, this method could not be applied due to FlorML's increased complexity and the sheer size of the documents (often several hundreds of pages).

Automation. — Therefore, it was decided to use the programming language Perl (http://www.perl.org/; v.5.12.3) to automate major parts of the mark-up process. Perl is a scriptable programming language that is both flexible and powerful. It is easy to learn and has advanced text-matching capabilities called regular expressions. These can be used to match specific text patterns and replace them with different text or insert additional text (see http://perldoc.perl.org/perlrequick.html for a primer). All of the Perl scripts were written in Notepad++ 5.9.3 (http://notepad-plus-plus.org/). The scripts use Unicode encoding (http://www.unicode.org/) to ensure they have no issues with any special symbols present in the treatments. The scripts will be made available on GitHub (see above).

The FlorML procedure for mark-up consists of three main steps: (1) Clean-up, (2) Automated mark-up, (3) File finalization. Several documents can be combined for mark-up. Figure 5 shows a schematic comparison of the workflow of our original mark-up procedure (Fig. 5A) and the one FlorML uses (Fig. 5B).

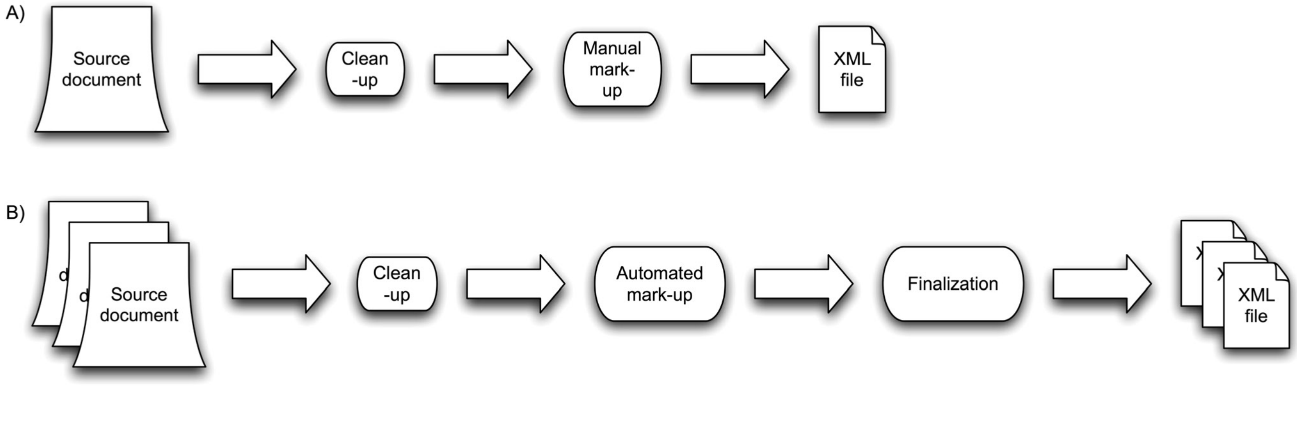


Fig. 5. Initial manual workflow **(A)** compared to current automated workflow **(B).** In the current automated workflow it is possible to combine multiple documents into one.

The workflow from a printed taxonomic work to an XML file starts by making the work available in digital form (see Introduction). Once digital versions of the legacy taxonomic treatments have been acquired, they are cleaned up manually in a word processor by deleting any unneeded text and removing all specific styles. While figures included in the document are removed, figure captions are kept in. The figures will be extracted from the original scans and converted to a format suitable for web use separately. Any text that interrupts sentences or paragraphs, such as figure captions, is moved to a better location. Indented keys are manually converted to linked polytomous keys, because the OCR process has major problems in properly recognizing the amount of white space in front of leads in indented keys, which is implicitly used for the key's hierarchy. It is suggested that any obvious typos or OCR errors in the text are fixed at this stage, especially if they are likely to interfere with the automated mark-up process.

The resulting file is saved as a plain text file with Unicode encoding. At this point two Perl scripts are run. The first one is a clean-up script that fixes common punctuation errors, removes excess white space including indentation, and standardises certain non-alphanumeric symbols. The second one fixes a few very common OCR errors that are hard to see, such as "P1." ("P-one-dot") instead of "Pl." ("P-lower case L-dot"). To enable one of the later scripts to make the distinction between each taxon, all taxa are then manually separated from each other using a single blank line.

The preparation process is summarized in Figure 6.

The automated mark-up with Perl scripts uses a simple yet effective approach. Each Perl script uses regular expressions to search for certain reoccurring patterns within a taxonomic work and inserts XML elements where appropriate. The script order is determined by analysing the structure of the taxonomic work, noting which text portions have specific text patterns. The scripts are ordered such that the text portions that are easiest to mark-up are marked up early on, after which subsequent scripts will ignore them unless needed and work on different text portions. Each script is specialized in a certain task, marking up only specific text portions and ignoring any text it does not need to change, whether previously marked up or not. In general, mark-up is first added to the larger structural text portions, before the remaining text in those text portions is atomised. Scripts that atomise contents are usually run towards the end of the mark-up process.

The first scripts that are always run insert the XML elements describing the basic structure of the taxonomic work, such as where the document begins and ends, and the basic mark-up for each taxon treatment. This is followed by a script that adds as much mark-up as possible to the keys, which use a very specific format that is not present elsewhere in the document. From there on, scripts are run in a predefined order to add mark-up to the rest of the document.

Although the Perl scripts for each Flora may perform mark-up of the same type of text portion, they need to be individually tailored to the actual contents of each taxonomic work. Likewise, as the script order is dependent on the document structure, different taxonomic works will have different script orders. Table 1 shows the order of scripts and executed tasks for both *Flora Malesiana* and *Flore du Gabon*, together with any required manual tasks. *Flore du Gabon* requires more manual work due to its loose document structure.

Each Perl script processes the text file line-by-line. When a text match occurs, the required XML elements are either prepended or suffixed to the text string, or the text string is split up into smaller parts so the XML elements can be inserted in between the various parts of the text string (Fig. 7A). Some scripts will ignore previously marked up text portions by using negative matches where the script looks for a text portion that does not contain a specific pattern (Fig. 7B). This approach is combined with conditional loops, where one specific action is taken when a certain pattern is present and another when another pattern is present (Fig. 7C). This makes it possible to, for example, mark up name types and specimen types using the same script, even though each requires specific mark-up. When atomizing nomenclatural data or literature references, the scripts first attempt to match very specific patterns, followed by increasingly broader ones (Fig. 7D). Perl's look-around assertions, which look for specific patterns preceding or following other patterns without actually using those specific patterns in the text replacement, are very useful for more advanced pattern matching.

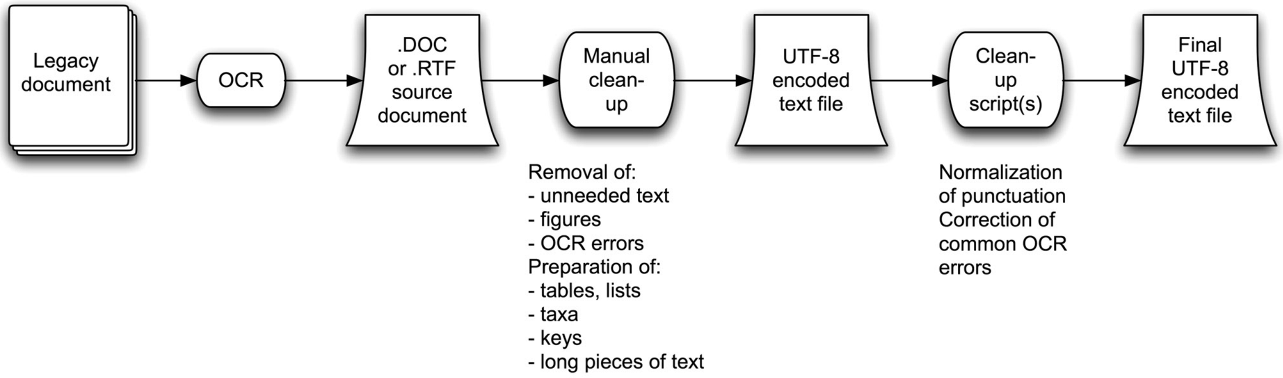


Fig. 6. Summary of legacy taxonomic treatment preparation.

Useful patterns for text matching are found in several parts of a taxonomic treatment. Some of the simplest patterns like headings and figure or table captions are reoccurring. Linked polytomous keys have a very structured and unique format that can be matched with relatively simple regular expressions. Although taxonomic descriptions make use of very variable character descriptors, the punctuation used to structure descriptions is very consistent. This also applies to large lists of literature references and citations. In nomenclature, similar possibilities can be exploited to recognise homotypic synonyms and heterotypic synonyms, while combining the punctuation with white space and commonly found naming schemes makes it possible to fully atomise most names. For example, a subspecies generally has a name that consists of a genus name, with a capitalized first letter, followed by a species name, an indication of subspecies, the subspecies name, and author information (Fig. 8A). The atomisation of literature references works in a comparable manner (Fig. 8B). The Perl scripts are written in such a way that no text aside from some punctuation and white space can be lost.

Most of the Perl scripts consist of long lists of regular expressions covering all possible options for a certain task. When a new option is discovered it is manually added to the Perl script and tested. The regular expressions used in each script are ordered in such a way that they do not interfere with each other. We suggest that it might be worthwhile to occasionally check whether certain scripts have correctly marked up all of their target contents just after running them to help discover previously unused patterns. Such a check can also be performed when there are reasons to believe that certain scripts may fail, such as a bad OCR.

Finalizing the XML file. — The resulting XML file will likely not be valid XML, where all opening elements are exactly matched by their accompanying closing elements. There are two main reasons for this. First, the scripts aim at inserting as much XML as practically possible, and not at creating well-formed XML. Second, it is possible that the Perl scripts failed to match some text that did not fit into any of the defined patterns or only partially matched text. This may cause some of the treatment text to end up in places where it is not allowed to be according to the FlorML schema or the insertion of an incorrect number of XML elements, resulting in validation errors.

Even after the XML file successfully validates errors still may be present, because text can be misidentified as being something else than it actually is by a regular expression. This is a problem that often occurs in text with formatting inconsistencies (mostly in literature references or citations).

Starting off with a plain text file prior to mark-up makes it impossible to add mark-up to text that is printed in bold or italics with a specific purpose or text that is sub- or superscripted, for example in chemical formulas. The scripts cannot recognise such advanced contents, because the scripts are not able to intelligently understand the text. Such text should be marked up manually at this stage, as should any other text for which mark-up is very hard to automate, such as habitat text strings. Metadata is also marked up manually.

Due to all of these reasons, XML files should be thoroughly checked against the printed original before they are ready for further processing.

Table 1. Script order and executed tasks. Similar scripts are marked using letters to clarify the different order.

Order	Task	Order	Task
1	clean-up (A)	1	clean-up (A)
2	fixing OCR errors (B)	2	fixing OCR errors (B)
M1	empty lines between taxa (C)	M1	empty lines between taxa (C)
3	parent document structure elements (D)	3	parent document structure elements (D)
4	taxon elements (E)	4	taxon elements (E)
5	keys (F)	5	keys (F)
M2	tables, lists and line breaks (G)	6	figures (J)
6	features excluding descriptions (H)	7	footnotes (I)
7	footnotes (I)	M2	tables, lists and line breaks (G)
8	figures (J)	8	feature basics (H – partial)
9	descriptions (K)	9	nomenclature basics (N)
10	description atomisation (L)	M3	remaining elements for features (H – partial)
11	taxon-specific headings (M)	10	nomenclature, literature references, descriptions atomisation (L, O, P)
12	nomenclature basics (N)	11	special symbols (R)
13	nomenclature atomisation (O)		
14	literature references basics and atomisation (P)		
15	author comments (Q)		
16	special symbols (R)		

M, manual task.

Figure 9 shows a fragment of text going through the various mark-up steps: (A) original text, (B) cleaned up text, (C) text during mark-up process, (D) final XML version.

Further processing. — Once the XML file has been finalized, it can be imported into the EDIT CDM database system using a special import script that matches every bit of XML annotated contents to the corresponding field in the database.

Another option is to archive the marked up taxonomic treatments in an online repository. Using XSLT (eXtensible Stylesheet Language Transformations) these can then be converted into the format (e.g., PDF) requested by the user on the fly.

■ RESULTS

***Flora Malesiana*. —** Nine volumes of *Flora Malesiana*, *Flora Malesiana* Series I volumes 14–20 and Series II volumes 2 and 3, have been marked up using the method described above. This consists of 3254 pages of printed treatments, covering around 2560 taxa belonging to 21 families. Prior to this, volume 11 part 3, volumes 12 and 13, covering a total of 1575 pages, 1605 taxa and 13 families had been marked up entirely manually. All these volumes are available through the *Flora Malesiana* Data portal at http://dev.e-taxonomy.eu/dataportal/flora-malesiana/.

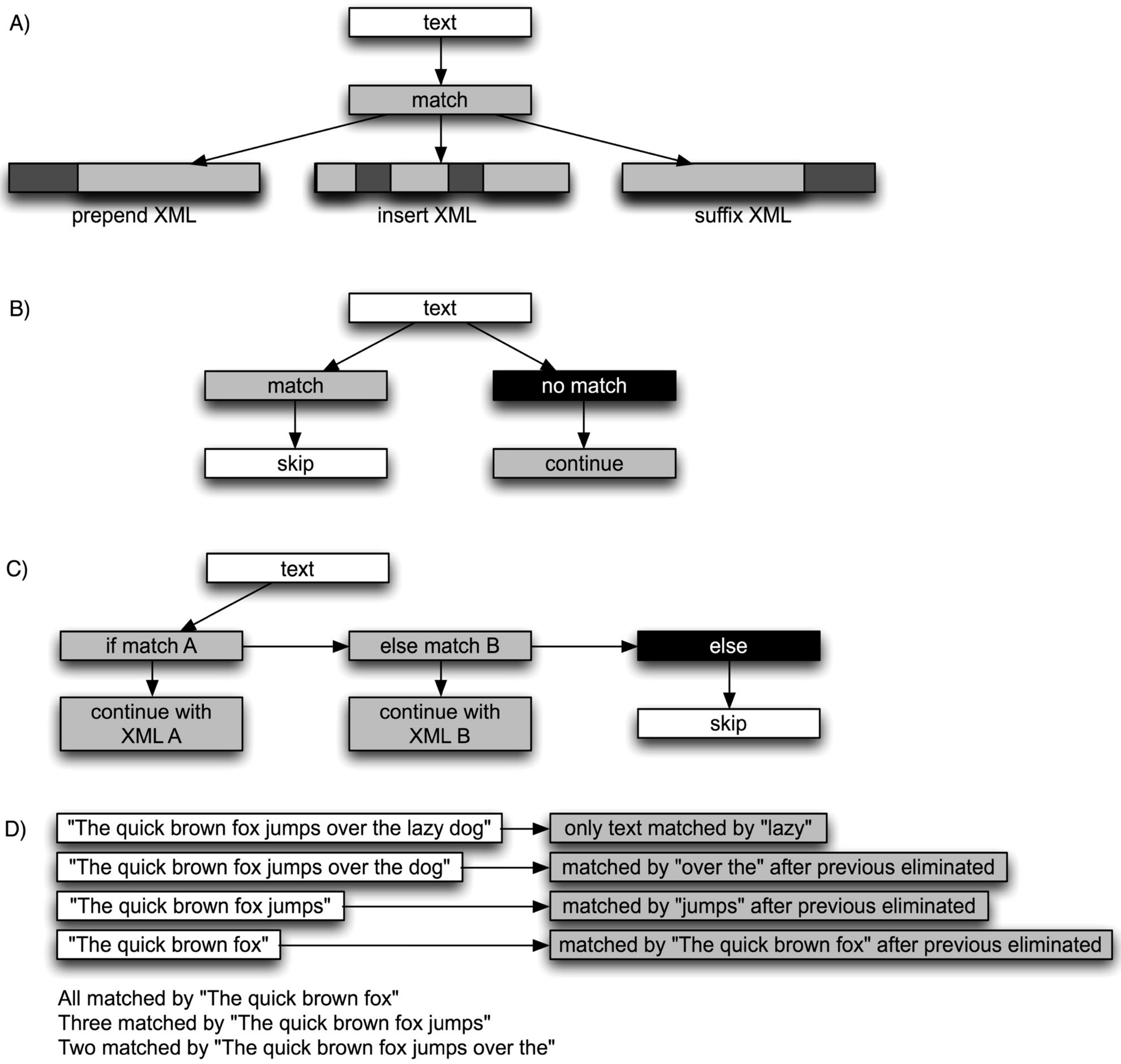

A)
text
match
prepend XML
insert XML
suffix XML
B)
text
match
no match
skip
continue
C)
text
if match A
else match B
else
continue with XML A
continue with XML B
skip
D)
"The quick brown fox jumps over the lazy dog"
only text matched by "lazy"
"The quick brown fox jumps over the dog"
matched by "over the" after previous eliminated
"The quick brown fox jumps"
matched by "jumps" after previous eliminated
"The quick brown fox"
matched by "The quick brown fox" after previous eliminated
All matched by "The quick brown fox"
Three matched by "The quick brown fox jumps"
Two matched by "The quick brown fox jumps over the"

Fig. 7. Script design choices. **A,** XML (dark grey parts) can be added in three different ways to a text fragment; **B,** negative matching; **C,** conditional looping; **D,** first match specific patterns, then increasingly broader ones.

***Flore du Gabon.* —** Volumes 1 to 22, 5bis, 27, 28 and 30 of *Flore du Gabon* have been marked up using the method described above, covering 4492 pages of printed treatments, with a total of about 3066 taxa belonging to 71 families. They are (slated to be made) available through the *Flore du Gabon* Data portal at http://dev.e-taxonomy.eu/dataportal/flore-gabon/.

Script development. — The initial development of the scripts for *Flora Malesiana* was combined with learning Perl (starting with no knowledge). It took about three months before the scripts were sufficiently bug-free and ready for use. Creating production-level scripts for *Flore du Gabon* took less than two months, by making some judicial reuse of previously developed scripts and avoiding earlier pitfalls. Preliminary results for the mark-up of Naturalis Biodiversity Center's third large Flora, the *Flora of the Guianas*, indicate that further reuse of previously developed scripts may reduce the initial development time for that Flora to as little as a single week, as this Flora uses a combination of formats found in the other two Floras.

Script use. — Use of the Perl scripts significantly improved both processing time and mark-up accuracy. For example, the mark-up process for *Flora Malesiana* Volume 13 (454 pages), which was performed mostly manually with help of Microsoft Word's find and replace function, took almost exactly four months. One of the first volumes to be marked up using non-prototype Perl scripts, *Flora Malesiana* Volume 14 (634 pages), took a mere 1.5 months, despite using a more advanced and complicated version of the FlorML XML schema compared to volume 13. These figures include text preparation prior to mark-up and post-processing of the resulting XML files.

Further improvements in processing time can be made by pasting similarly formatted legacy taxonomic publications end to end into one single text file and running the Perl scripts on this file instead of each of the individual files. Using this method, *Flora Malesiana* Volume 17 Parts 1 and 2 (884 pages combined) were marked up in less than a month. This method has become the default method for marking up *Flore du Gabon*, where the latest batch of volumes (vols. 12–20, over 2200 pages total) was marked up in 1.6 months. However, the success rate of this improvement is dependent on the absence of divergent structural elements in the text that could interfere with the Perl scripts.

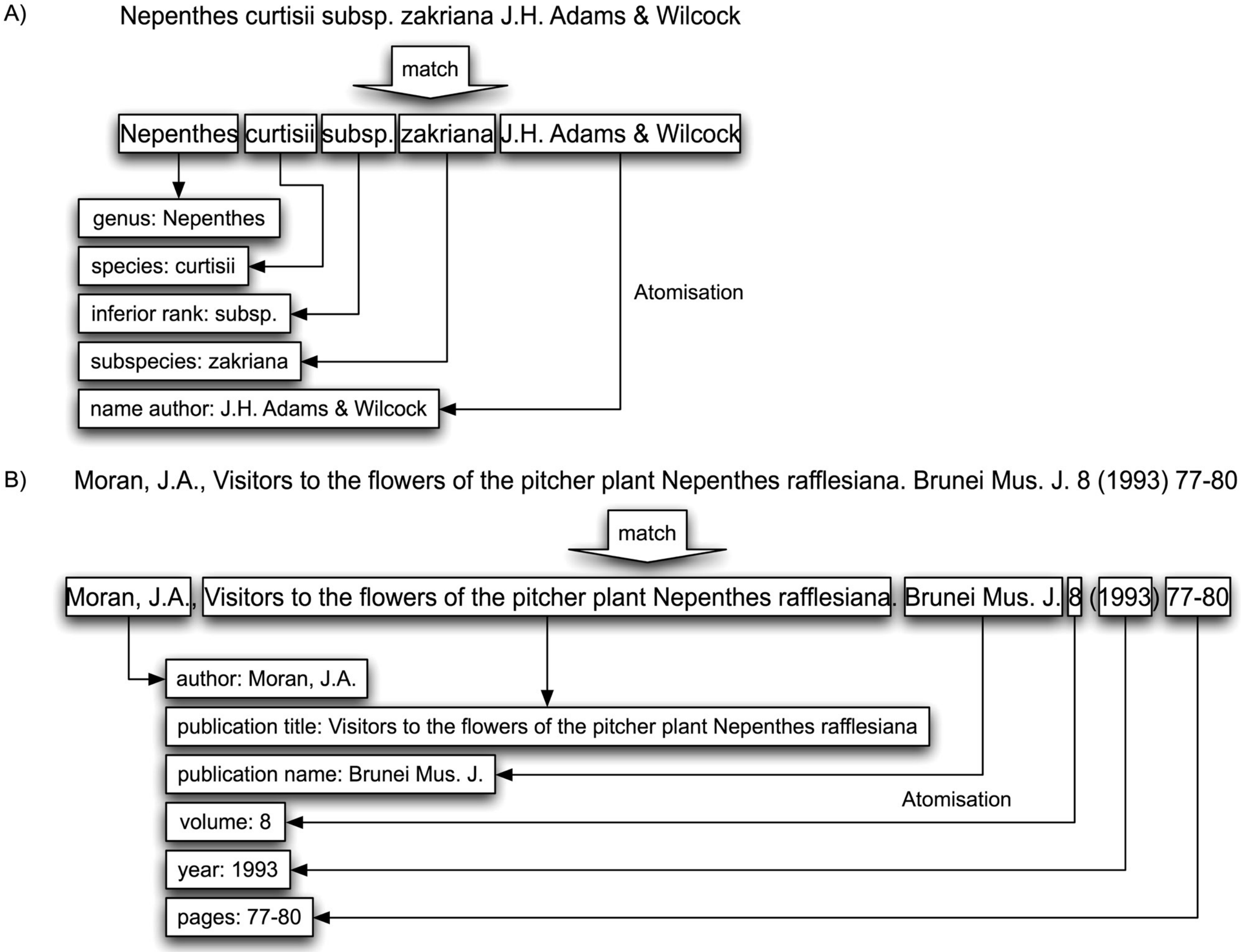
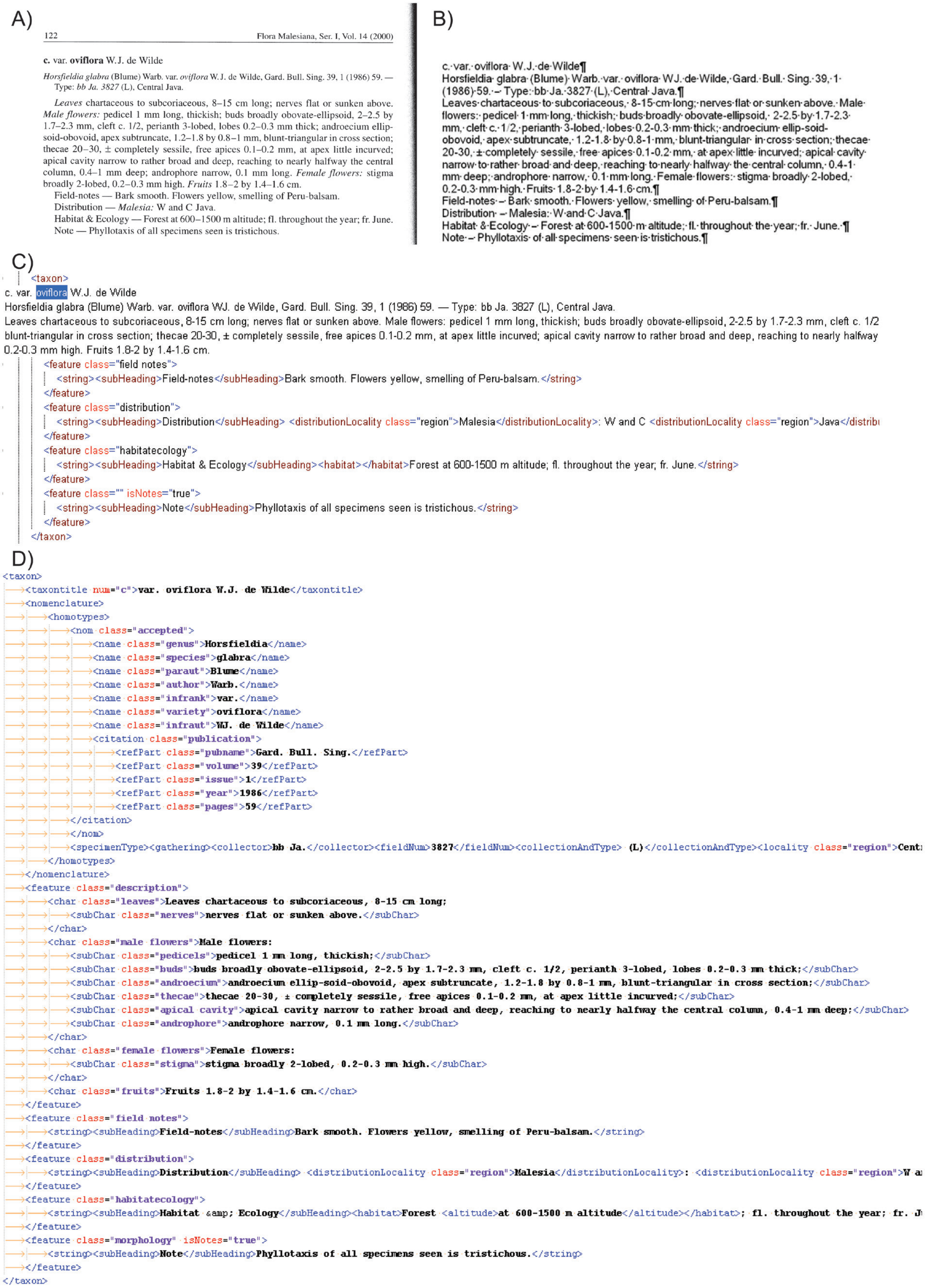

A)
Nepenthes curtisii subsp. zakriana J.H. Adams & Wilcock
match
Nepenthes
curtisii
subsp.
zakriana
J.H. Adams & Wilcock
genus: Nepenthes
species: curtisii
inferior rank: subsp.
subspecies: zakriana
name author: J.H. Adams & Wilcock
Atomisation
B)
Moran, J.A., Visitors to the flowers of the pitcher plant Nepenthes rafflesiana. Brunei Mus. J. 8 (1993) 77-80
match
Moran, J.A.
Visitors to the flowers of the pitcher plant Nepenthes rafflesiana
Brunei Mus. J.
8
1993
77-80
author: Moran, J.A.
publication title: Visitors to the flowers of the pitcher plant Nepenthes rafflesiana
publication name: Brunei Mus. J.
volume: 8
year: 1993
pages: 77-80
Atomisation

Fig. 8. Examples of using alphanumeric characters in combination with white space and punctuation to match and atomise text. **A,** atomisation of a subspecies; **B,** atomisation of a literature reference.



A)

122 Flora Malesiana, Ser. I, Vol. 14 (2000)

c. var. **oviflora** W.J. de Wilde

Horsfieldia glabra (Blume) Warb. var. *oviflora* W.J. de Wilde, Gard. Bull. Sing. 39, 1 (1986) 59. — Type: *bb Ja. 3827* (L), Central Java.

Leaves chartaceous to subcoriaceous, 8–15 cm long; nerves flat or sunken above. *Male flowers:* pedicel 1 mm long, thickish; buds broadly obovate-ellipsoid, 2–2.5 by 1.7–2.3 mm, cleft c. 1/2, perianth 3-lobed, lobes 0.2–0.3 mm thick; androecium ellipsoid-obovoid, apex subtruncate, 1.2–1.8 by 0.8–1 mm, blunt-triangular in cross section; thecae 20–30, ± completely sessile, free apices 0.1–0.2 mm, at apex little incurved; apical cavity narrow to rather broad and deep, reaching to nearly halfway the central column, 0.4–1 mm deep; androphore narrow, 0.1 mm long. *Female flowers:* stigma broadly 2-lobed, 0.2–0.3 mm high. *Fruits* 1.8–2 by 1.4–1.6 cm.

Field-notes — Bark smooth. Flowers yellow, smelling of Peru-balsam.

Distribution — *Malesia:* W and C Java.

Habitat & Ecology — Forest at 600–1500 m altitude; fl. throughout the year; fr. June.

Note — Phyllotaxis of all specimens seen is tristichous.

B)

c.·var.·oviflora·W.J.·de·Wilde¶
Horsfieldia·glabra·(Blume)·Warb.·var.·oviflora·WJ.·de·Wilde,·Gard.·Bull.·Sing.·39,·1·(1986)·59.·–·Type:·bb·Ja.·3827·(L),·Central·Java.¶
Leaves·chartaceous·to·subcoriaceous,·8-15·cm·long;·nerves·flat·or·sunken·above.·Male·flowers:·pedicel·1·mm·long,·thickish;·buds·broadly·obovate-ellipsoid,·2-2.5·by·1.7-2.3·mm,·cleft·c.·1/2,·perianth·3-lobed,·lobes·0.2-0.3·mm·thick;·androecium·ellip-soid-obovoid,·apex·subtruncate,·1.2-1.8·by·0.8-1·mm,·blunt-triangular·in·cross·section;·thecae·20-30,·±·completely·sessile,·free·apices·0.1-0.2·mm,·at·apex·little·incurved;·apical·cavity·narrow·to·rather·broad·and·deep,·reaching·to·nearly·halfway·the·central·column,·0.4-1·mm·deep;·androphore·narrow,·0.1·mm·long.·Female·flowers:·stigma·broadly·2-lobed,·0.2-0.3·mm·high.·Fruits·1.8-2·by·1.4-1.6·cm.¶
Field-notes·–·Bark·smooth.·Flowers·yellow,·smelling·of·Peru-balsam.¶
Distribution·–·Malesia:·W·and·C·Java.¶
Habitat·&·Ecology·–·Forest·at·600-1500·m·altitude;·fl.·throughout·the·year;·fr.·June.·¶
Note·–·Phyllotaxis·of·all·specimens·seen·is·tristichous.¶

C)

```
    <taxon>
c. var. oviflora W.J. de Wilde
Horsfieldia glabra (Blume) Warb. var. oviflora WJ. de Wilde, Gard. Bull. Sing. 39, 1 (1986) 59. — Type: bb Ja. 3827 (L), Central Java.
Leaves chartaceous to subcoriaceous, 8-15 cm long; nerves flat or sunken above. Male flowers: pedicel 1 mm long, thickish; buds broadly obovate-ellipsoid, 2-2.5 by 1.7-2.3 mm, cleft c. 1/2
blunt-triangular in cross section; thecae 20-30, ± completely sessile, free apices 0.1-0.2 mm, at apex little incurved; apical cavity narrow to rather broad and deep, reaching to nearly halfway
0.2-0.3 mm high. Fruits 1.8-2 by 1.4-1.6 cm.
        <feature class="field notes">
          <string><subHeading>Field-notes</subHeading>Bark smooth. Flowers yellow, smelling of Peru-balsam.</string>
        </feature>
        <feature class="distribution">
          <string><subHeading>Distribution</subHeading> <distributionLocality class="region">Malesia</distributionLocality>: W and C <distributionLocality class="region">Java</distribu
        </feature>
        <feature class="habitatecology">
          <string><subHeading>Habitat & Ecology</subHeading><habitat></habitat>Forest at 600-1500 m altitude; fl. throughout the year; fr. June.</string>
        </feature>
        <feature class="" isNotes="true">
          <string><subHeading>Note</subHeading>Phyllotaxis of all specimens seen is tristichous.</string>
        </feature>
    </taxon>
```

D)

```
<taxon>
    <taxontitle num="c">var. oviflora W.J. de Wilde</taxontitle>
    <nomenclature>
        <homotypes>
            <nom class="accepted">
                <name class="genus">Horsfieldia</name>
                <name class="species">glabra</name>
                <name class="paraut">Blume</name>
                <name class="author">Warb.</name>
                <name class="infrank">var.</name>
                <name class="variety">oviflora</name>
                <name class="infraut">WJ. de Wilde</name>
                <citation class="publication">
                    <refPart class="pubname">Gard. Bull. Sing.</refPart>
                    <refPart class="volume">39</refPart>
                    <refPart class="issue">1</refPart>
                    <refPart class="year">1986</refPart>
                    <refPart class="pages">59</refPart>
            </citation>
            </nom>
            <specimenType><gathering><collector>bb Ja.</collector><fieldNum>3827</fieldNum><collectionAndType> (L)</collectionAndType><locality class="region">Cent
        </homotypes>
    </nomenclature>
    <feature class="description">
        <char class="leaves">Leaves chartaceous to subcoriaceous, 8-15 cm long;
            <subChar class="nerves">nerves flat or sunken above.</subChar>
        </char>
        <char class="male flowers">Male flowers:
            <subChar class="pedicels">pedicel 1 mm long, thickish;</subChar>
            <subChar class="buds">buds broadly obovate-ellipsoid, 2-2.5 by 1.7-2.3 mm, cleft c. 1/2, perianth 3-lobed, lobes 0.2-0.3 mm thick;</subChar>
            <subChar class="androecium">androecium ellip-soid-obovoid, apex subtruncate, 1.2-1.8 by 0.8-1 mm, blunt-triangular in cross section;</subChar>
            <subChar class="thecae">thecae 20-30, ± completely sessile, free apices 0.1-0.2 mm, at apex little incurved;</subChar>
            <subChar class="apical cavity">apical cavity narrow to rather broad and deep, reaching to nearly halfway the central column, 0.4-1 mm deep;</subChar>
            <subChar class="androphore">androphore narrow, 0.1 mm long.</subChar>
        </char>
        <char class="female flowers">Female flowers:
            <subChar class="stigma">stigma broadly 2-lobed, 0.2-0.3 mm high.</subChar>
        </char>
        <char class="fruits">Fruits 1.8-2 by 1.4-1.6 cm.</char>
    </feature>
    <feature class="field notes">
        <string><subHeading>Field-notes</subHeading>Bark smooth. Flowers yellow, smelling of Peru-balsam.</string>
    </feature>
    <feature class="distribution">
        <string><subHeading>Distribution</subHeading> <distributionLocality class="region">Malesia</distributionLocality>: <distributionLocality class="region">W a
    </feature>
    <feature class="habitatecology">
        <string><subHeading>Habitat &amp; Ecology</subHeading><habitat>Forest <altitude>at 600-1500 m altitude</altitude></habitat>; fl. throughout the year; fr. J
    </feature>
    <feature class="morphology" isNotes="true">
        <string><subHeading>Note</subHeading>Phyllotaxis of all specimens seen is tristichous.</string>
    </feature>
</taxon>
```

Fig. 9. Mark-up steps, using part of a page of *Flora Malesiana* Vol. 14 as an example. **A,** original text as it appears in the printed volume; **B,** cleaned up text in Microsoft Word; **C,** XML version halfway through automated mark-up process; **D,** final XML version.

The Perl scripts allow for up to 80% of the mark-up to be added automatically, going up to 90%–95% for volumes with few divergent elements. Table 2 gives some typical estimated quantitative metrics for the more advanced scripts for each of the Floras. The results can be adequately explained by recalling that the scripts use pure text pattern matching for inserting XML instead of a probabilistic method. False positives occur whenever there is an accidental match with non-target text and are generally rare, which is reflected in the high precision and low error ratio scores. With some scripts, false positives are impossible due to verbatim text matching. False negatives are caused either by formats or keywords that are absent from the scripts, or by OCR errors interfering with proper mark-up. The recall score can be seen as a measure for how well a script is matched to the texts it is used to process. However, it is impossible to say whether a relatively low score is caused by a lack of corresponding regular expressions in the script or because most unmatched text has fairly unique formats that are encountered only once and therefore are not added to the script. Based on our proofreading experiences the second option seems to be most likely.

Issues encountered prior to mark-up. — Table 3 lists the most common problems found prior to XML mark-up and the solutions we used.

Typos are fairly rare compared to OCR errors. The most common sources of OCR errors are symbols, punctuation and white space that are misidentified during the OCR process, but stains and other dirt on the pages may also be interpreted as symbols by the OCR software. Spotting such errors requires carefully proofreading the whole taxonomic work, which takes a considerable amount of time and will not be fully successful due to the similarity of certain symbols, such as lower case

Table 2. Estimated quantitative performance metrics for advanced Perl scripts.

Script	FP (%)	FN (%)	E	P	R	F1
Flora Malesiana						
keys	0.046	0.898	0.009	1.000	0.991	0.995
features (excl. descriptions)	0.000	12.798	0.128	1.000	0.887	0.940
footnotes	0.000	31.707	0.317	1.000	0.759	0.863
figures	0.000	17.731	0.177	1.000	0.849	0.919
descriptions	0.138	0.138	0.003	0.999	0.999	0.999
description atomisation	1.195	0.005	0.012	0.988	1.000	0.994
taxon-specific headings	0.496	1.309	0.018	0.995	0.987	0.991
nomenclature basics	1.482	2.762	0.042	0.985	0.973	0.979
nomenclature atomisation	2.612	23.671	0.263	0.975	0.809	0.884
literature references basics and atomisation	1.763	13.503	0.153	0.983	0.881	0.929
Flore du Gabon						
keys	0.050	7.941	0.080	1.000	0.926	0.962
figures	0.582	11.270	0.119	0.994	0.899	0.944
footnotes	1.064	37.234	0.383	0.989	0.729	0.839
features basics	1.799	32.663	0.345	0.982	0.754	0.853
nomenclature basics	0.030	14.801	0.148	1.000	0.871	0.931
nomenclature, literature references, descriptions atomisation	1.664	9.082	0.107	0.984	0.917	0.949

FP, false positives; FN, false negatives; E, number of errors per correctly inserted XML element; P, precision; R, recall; F1, script accuracy based on regular expressions used in script.

Table 3. Most common problems encountered in taxonomic treatments prior to XML mark-up and our solutions.

Problem	Solution
Typo in original document	Fix manually if interfering with mark-up; else ignore
OCR error: common alphanumeric symbol mix-up	Fix manually or using Perl script, if required add error to script
OCR error: symbol or character not recognised as text	Manually replace by correct symbol or character during initial clean-up phase
Obvious errors in punctuation, brackets, and dashes	Run clean-up Perl script, if required add erroneous text to clean-up script
Unrequired white space	Run clean-up Perl script
Indented key	Manually convert to regular polytomous key
Text in improper location that will interfere with mark-up process	Manually move text to better location

"1" and the number "1". In some cases, specific symbols are not recognized as text but as graphics by the OCR software. These are always fixed.

Unrequired white space consists of multiple subsequent spaces or tabs. White space is also used to provide the hierarchy of indented keys in many legacy taxonomic treatments. Each indent before a lead in an indented key uses a specific amount of white space to indicate its relation with the previous lead. Unfortunately, this kind of white space generally does not survive the OCR process unscathed. Because manually fixing this proved very time-consuming, we instead converted indented keys to linked polytomous keys by hand, and use the script we developed for polytomous keys to apply mark-up.

Issues encountered during the mark-up process. — Various issues were encountered during the mark-up process. Table 4 lists the most common problems found during the XML file finalization process, their reason, and the long-term solution. Long term solutions are only implemented when an issue can be expected to occur repeatedly. The short-term solution is to fix the error directly in the XML file, and in some cases this is the only solution practically available.

The errors encountered after running the Perl scripts for automated mark-up can be categorised as (1) text not matching or mismatching a regular expression and (2) text misidentification. These are caused by a combination of certain text patterns and the regular expressions used in the scripts. In general, the solution is to fix the error manually in the file being finalized and then add the text pattern to the corresponding script.

However, in the case of text misidentification text pattern recognition may not be the only problem, but the actual text may be ambiguous. An example of this is the text "25, Plate 3" following a year in a literature reference or citation. This can be interpreted as information indicating there is a figure "Plate 3 on page 25", but also as "Page 25" mentioning the name and a figure "Plate 3" on another, sometimes unnumbered page elsewhere. Unfortunately, it is impossible to determine which of the options is correct without physically checking the actual publication. Another example is that authors sometimes use shorthand in parts of a treatment, by merging repetitive information or leaving out information that was explicitly given earlier.

Issues with legacy taxonomic work contents. — The two previous examples are part of a third type of issue that we encountered: vague contents. This has little to do with the actual XML schema design and mark-up, but relates directly to the contents of the work. Quite regularly, the author assumes the reader has some specific prior (implicit) knowledge that facilitates interpretation of the contents, or that the reader can deduct obviously missing contents based on their own knowledge (see also Hagedorn, 2007).

Such implicit knowledge can sometimes be mentioned beforehand in another part of the publication, but more often it is not present anywhere in the taxonomic work. However, as the Perl scripts only look for text patterns to match and do not actually understand the text or context, this vague contents can only be properly marked up manually.

Table 4. Most common problems encountered in taxonomic treatments after XML mark-up, their reason, and the long-term solution.

Problem	Reason	Long-term solution
(Sub-)Heading not marked up	(Sub-)Heading text option missing from script	Add (sub-)heading text to script if possible
Description not marked up	No description-specific keyword present	Discover description-specific keyword and add to script
Key lead not marked up	Format not recognized, no match	Add new format to script if possible
Taxonomic name, type, specimen, or reference not atomised	Format not recognized, no match	Add new format to script if possible
Taxonomic name, type, specimen, or reference partially atomised	Format not recognized, partial match with existing regular expression	Add new format to script if possible
Name, type specimen, or reference parts misidentified	Format not recognized, unexpected partial match with existing regular expression	Add new format to script if possible
(Sub-)Character not or misidentified	(Sub-)Character text option missing from script or unexpected partial match with existing regular expression	Fix manually in XML file, add new format to script if possible
Distribution locality not or misidentified	Text not recognized, no or unexpected partial match with existing regular expression	Add new format to script if possible
Figure caption or reference to figure not marked up	Format not recognized, no match	Add new format to script if possible
Element wrongly inserted in unexpected location	Unexpected partial match with existing regular expression	Fix manually in XML file

A similar problem occurs in indications of character dimensions. Whether a measurement is a length, width, height, or diameter is not always indicated, and should then be derived from the definition of the character the measurement applies to. When two measurements are given for a character, e.g., "leaves 15 × 10 mm", which of the two measurements is "length" and which is "width" is left to the reader. It likely indicates the leaves are 15 mm long and 10 mm wide, although it might also mean they are 15 mm wide and 10 mm long, depending on the preferred order for measurements of the treatment author. A simple way of bypassing this vagueness would be to understand this as "measurements = 10 × 15 mm", instead of attempting to atomise it into its two components length and width. This can still be marked up using a simple regular expression. However, consider that the same description contains a line reading "stem 10 × 15 mm". This could mean a stem with a width of 10 mm and a length of 15 mm (or the opposite). It could also mean that the cross-section of the stem is 10 by 15 mm. Which of the two is correct is left to the reader, who will likely conclude that the second option probably is the correct one. Based on this choice, a script can be written that explicitly states that measurements for a stem are the cross-section. However, using regular expressions only it is impossible to write a script that actually deduces the choice to make.

A somewhat more complicated issue is present in the definitions of characters themselves. More often than not, no character definitions are offered by the author(s) of a legacy taxonomic work. Determining how the writer of a taxonomic treatment interpreted various characters by comparing their descriptions to actual specimens is therefore an important step in taxonomic revisions. It is impractical to do this while marking up a legacy taxonomic work, because it is very time-consuming. Furthermore, character definitions may actually differ depending on the taxonomic group, even though they use the same term in taxon descriptions. Differences in interpretation of characters by different scientists means different definitions exist. Because a scientist's interpretation and understanding of characters may change during their lifetime, character definition differences are also present within the work of a single author. Some scientists may be unsure of how to precisely define a character, which may cause considerable variation of terminology use even within a small taxonomic work (Lydon & al., 2003; Cui & Heidorn, 2007; Hagedorn, 2007; Seltmann & al., 2013; Thessen & al., 2012b; P. Hovenkamp, R. Sluys, pers. comm.).

We have mentioned that FlorML supports over 900 different taxonomic characters. Some of these may be synonyms, but it is not possible to determine this with certainty due to the issues discussed above.

■ DISCUSSION AND CONCLUSIONS

In this article we have described the development and deployment of FlorML, a new XML schema for marking up complicated legacy taxonomic works, such as extensive semi-monographic Floras, using a semi-structured approach. FlorML was designed by analysing the structure of these works, and identifying their types of contents and where these can occur in a taxonomic work. FlorML divides a taxonomic work into four different content types: metadata, taxonomic contents, non-taxonomic contents, and errata to previous volumes, each of which is subdivided further. The most important type, taxonomic contents, can be subdivided into keys, nomenclature, references and descriptions as well as a variety of other features. A further possible subdivision (atomisation) of certain contents is required for certain purposes, such as the creation of interactive multi-access keys. Compared to several other XML schemas in use (Weitzman & Lyal, 2004; Sautter & al., 2007; Penev & al., 2011), FlorML provides a much more detailed atomisation, to deal with the complex structure of taxonomic works.

Although we encountered problems with schema complexity early on during the development of FlorML, these were resolved by carefully analysing which parts of the XML schema could be simplified. The number of elements in the XML schema was reduced by using XML attributes that conferred to single XML elements the ability to deal with a multitude of similar contents. Many other schemas use separate elements for the subdivision of similar contents (Sautter & al., 2007; Cui, 2008a, b; Cui & al., 2010) instead of a single element with an attribute. Furthermore, elements were reused whenever possible. For the metadata an external XML schema, MODS, was used. The further development of FlorML mostly consisted of small additions to broaden the versatility of the XML schema. A larger change was expanding the character mark-up model to deal with descriptions having the same characters for both genders of flowers. FlorML will continue to be improved by the addition of support for more taxonomic characters as they become needed. This is an unavoidable step in XML schema development that was also noted by others (Cui, 2008a, b).

Improper clean-up of a taxonomic work will lead to complications during the automated mark-up process using FlorML. Incorrectly checked final XML files will either not validate or have improperly identified information at one or more points in the file. Kirkup & al. (2005) and Cui (2008a) also noted that the clean-up of taxonomic works prior to mark-up and the final verification of each XML file are steps that can not be avoided to get a satisfactory result, but are relatively time-consuming.

Scripts written in the scripting language Perl were highly effective to improve the accuracy, consistency and speed of the XML mark-up process using FlorML compared to fully manual mark-up. The number of corrections to be made during proofreading of the final XML file was also considerably reduced. Similar decreases in error rate due to automation were noted by Cui (2008b). The individual scripts make use of Perl's abilities to recognize and manipulate reoccurring text patterns to insert most of the XML. The text patterns used are often remarkably stable throughout a Flora, likely because many are formats enforced by either the editors of the taxonomic works or nomenclatural rules. Up until now acceleration factors of more than nine have been achieved compared to fully manual mark-up. Choosing a smart script order and excluding previously marked up sections in certain scripts is also very important for good results. Bundling similarly formatted legacy taxonomic publications prior to running the Perl scripts over all of them at once results in further gains in processing time. Further improvements to the success rate of the Perl scripts and the processing time will be seen by adding more regular expressions to each script to increase the text matching success rate.

In the future, we envisage deploying FlorML to mark up the third large Flora project of Naturalis Biodiversity Center, *Flora of the Guianas*. Preliminary testing using unmodified *Flora Malesiana* scripts on this Flora has revealed that the adaptability of the scripts greatly exceeds our expectations. Hopefully, other organisations thinking of transforming their written legacy taxonomic information to an e-version will also start using the FlorML schema. If opportunities present themselves, applying the same routines on faunistic literature will be attempted.

Although our results are generally very encouraging, we have also encountered a variety of issues complicating the XML mark-up process. Most issues are less related to the XML schema and XML mark-up process than they are to the contents of the taxonomic work in itself.

Some issues need to be addressed during clean-up prior to mark-up or the final check of the XML document. Because of our pragmatic approach to the mark-up process, we decided that having to move around or slightly modify small pieces of text prior to the automatic mark-up process to be able to successfully fit them into the XML schema in a small minority of cases was an acceptable alternative to a more complex XML schema to accommodate all possible options. It appeared most time-efficient to correct typos and OCR errors only when they are likely to interfere with the mark-up process, manually or with a script for the more common errors. One drawback is that not all such errors are corrected. Possible solutions for this issue would be to place the source documents on Wikisource or Mechanical Turk with PDFs or scans of the original work, and use crowdsourcing to proofread them and correct any typos and OCR errors, or train a neural network to spot errors (Thessen & al., 2012a; Thomer & al., 2012; G. Hagedorn, pers. comm.; R. Vos, pers. comm.).

Most issues related to automated mark-up and atomisation can be resolved by improving the Perl scripts to include more regular expressions. Additional regular expressions may only be part of the solution for misidentification of contents and wrongly inserted elements, especially if the problem is caused by vague or implicit contents. In some cases, creating lists of associated contents may provide a solution. For example, to properly mark up a publication's category (book, book part, journal, or something else), the script would need to include a list matching each publication with the corresponding category. However, compiling such lists may be rather time-consuming. They are only a viable option for concrete types of contents where associations can be precisely determined. Similar lists could also be created for taxonomic characters, but in many cases it is unfortunately very difficult to determine which characters truly are synonyms of each other. To be able to do this, extensive knowledge of the characters described is required, including knowledge of the specific morphology of each of the taxonomic groups described in a legacy taxonomic work (Hagedorn, 2007). Due to the large number of possible characters and the very variable nature of taxonomic groups this is a nearly impossible task for one person. For this reason we currently treat differently named characters as different characters.

Another solution may be a process called "machine learning". In machine learning a computer program uses algorithms involving probability and/or known patterns to determine what kind of contents it is looking at, aided in this task by training examples, the ability to discover new patterns and sometimes human assistance. Some promising research has been done on this subject, although it is limited to a small part (descriptions) of taxonomic works only. However, machine learning requires a fairly long development time to create the program and training examples for a single taxonomic work (Cui, 2008a; Cui & al 2010; Thessen & al., 2012a). Machine learning would likely yield better results with taxonomic works that have a loose structure, such as the early volumes of *Flore du Gabon*, because instead of using the structural elements of the work it depends on hidden patterns in the text itself. However, considering our current XML insertion rates using only regular expressions, it can be questioned whether machine learning would actually be a large improvement and worth the invested time.

Other issues need further action beyond the scope of XML schema design and mark-up. These include dealing with vague or implicit contents. This could be solved by applying ontologies that clearly define characters and their relationships (see also Lydon & al., 2003; Hagedorn, 2007; Thessen & Patterson, 2011), similar to how standardised characters are in use in several fields of biology, including those that deal with contents that is similar to that encountered in taxonomic descriptions, e.g., wood anatomy (e.g., IAWA Committee, 1989). Lessons can be drawn from such experiences with standardised character lists to increase the likelihood of having useful definitions. Crowd-sourcing, under specialist guidance, may be considered an option to obtain the information required for standardized ontologies, especially if this could lead to more generalized definitions. However, when definitions used in legacy taxonomic works are too variable, even ontologies can not resolve this problem. One solution might be to create a separate ontology to fit each and every legacy taxonomic work. Unfortunately, this method still does not take into account that character definitions may be variable even within a work. A more interesting option would be to conceptualize the problematic types of contents. Contents could then have a fairly generic general definition that encompasses the various taxon- and author-specific definitions. This leads to a model similar to the "character plus concept hierarchy" described by Hagedorn (2007). Such an approach unfortunately still requires a considerable amount of work by taxonomists, because each of the component definitions has to be determined.

An external factor complicating mark-up and especially atomisation is the difference in stakeholder expectations regarding the requirements of what should be atomised and how far atomisation should go. Opinions on how far atomisation has to go range from low-level atomisation going no further than whole plant structures to full atomisation of all details. The latter could enable automatic creation of advanced multi-entry keys. Furthermore, depending on the stakeholders' professional objectives the sections of a taxonomic work used are different. This means ideally all contents in a taxonomic work would have to be atomised, instead of being limited to the contents that we currently atomise, which is most suited for other taxonomists. Some stakeholders admit they would ideally want all the information atomised, but realize there are practical obstacles. Similar issues were noted by Thessen & Patterson (2011) and Morrison (2012). Even if those practical obstacles are resolved, full atomisation may not be possible due to practical and time constraints. These are not caused by the automated mark-up process itself, but by the time required to develop the scripts and check the final XML file. Some parts of a taxonomic work are highly structured and easy to automate for, but descriptive data often is vaguer in structure and terms, meaning that atomising them increases the probability that contents is encountered that is difficult to atomise. There is a point where expectations become too high compared to feasibility. A possible solution would be to limit atomisation to a certain level, and develop better search engines to search the data to extract the required information.

The current approach will be assessed and further developed within the framework of the EU-funded project pro-iBiosphere. Here the focus will be on both compatibility with other mark-up approaches and further atomisation of taxon descriptions to a degree that allows us to build identification keys. Compatibility with TaxonX will be examined as this format is actively used for marking-up data with the mark-up tool Golden Gate (http://plazi.org/?q=GoldenGATE) and creates similar results by using different approaches in terms of technology and algorithms. A combination of both approaches may improve the results. A combination may include either a common workflow in which both technologies are integrated or be a mere set of transformation tools to transform the results of one approach to the formats used by the other and thus make the results available to a larger set of consuming applications. Furthermore, the relationship to mark-up formats designed for publishing prospective data such as TaxPub (Catapano, 2010) will be further explored within pro-iBiosphere.

To conclude, we have been able to deploy far-reaching XML mark-up and atomisation of botanical legacy taxonomic works using FlorML, and managed to obtain significant improvements in accuracy and speed of mark-up by using the script language Perl. However, solutions will be needed to deal with the problem of vague and implicit contents to enable the functional exchange of descriptive biodiversity data on the web.

Meanwhile, we will continue to improve our deployment of FlorML, aiming at the complete digitalisation of *Flora Malesiana* and *Flore du Gabon*.

■ ACKNOWLEDGEMENTS

We would like to thank Don Kirkup for helping us get this project started by kindly letting us base our initial XML mark-up on his work. We also would like to thank everyone at Naturalis Biodiversity Center and beyond, who has weighed in with advice and suggestions to improve the quality of our work during the development of the FlorML XML schema and the writing of this article. Finally, the first author would like to personally thank J.F. Veldkamp of Naturalis Biodiversity Center, without whose taxonomic guidance the present work could not have been completed. This project was subsidized in part by the EU project "pro-iBiosphere" (Grant agreement 312848).

■ LITERATURE CITED

Barber, A., Lafferty, D. & Landrum, L.R. 2013. The SALIX Method: A semi-automated workflow for herbarium specimen digitization. *Taxon* 62: 581–590. http://dx.doi.org/10.12705/623.16

Berendsohn, W.G. 2010. Devising the EDIT platform for cybertaxonomy. Pp. 1–6 in: Nimis, P.L. & Vignes-Lebbe, R. (eds.), *Tools for identifying biodiversity: Progress and problems*. Trieste: Edizioni Università di Trieste.

Berendsohn, W.G., Güntsch, A., Hoffmann, N., Kohlbecker, A., Luther, K. & Müller, A. 2011. Biodiversity information platforms: From standards to interoperability. *ZooKeys* 150: 71–87. http://dx.doi.org/10.3897/zookeys.150.2166

Biron, P.V. & Malhotra, A. 28 Oct 2004. XML schema part 2: Datatypes second edition. W3C recommendation. http://www.w3.org/TR/2004/REC-xmlschema-2-20041028/datatypes.html (accessed 26 Sep 2012).

Catapano T. 2010. TaxPub: An extension of the NLM/NCBI journal publishing DTD for taxonomic descriptions. In: *Journal Article Tag Suite Conference (JATS-Con)*, Proceedings 2010. Bethesda: National Center for Biotechnology Information. http://www.ncbi.nlm.nih.gov/books/NBK47081/ (accessed 27 May 2013).

Cleal, C.J. & Thomas, B.A. 2010a. Botanical nomenclature and fossil plants. *Taxon* 59: 261–268.

Cleal, C.J. & Thomas, B.A. 2010b. Proposals to modify the provisions in the *Code* for naming plant fossils. *Taxon* 59: 303–313.

Cleal, C.J. 2011. New regulations affecting taxonomic nomenclature of plant fossils. *Linn. Soc. Palaeobot. Specialist Group: Palaeobot. Group Newslett.* 25: 12.

Corney, D.P.A., Clark, J.Y., Tang, H.L. & Wilkin, P. 2012. Automatic extraction of leaf characters from herbarium specimens. *Taxon* 61: 231–244.

Cui, H. 2008a. Approaches to semantic markup for natural heritage literature. *Proceedings of the iConference 2008*. http://ischools.org/conference08/pc/PA5-2_iconf08.doc (accessed 21 Sep 2012).

Cui, H. 2008b. Converting taxonomic description to new digital formats. *Biodivers. Informatics* 5: 20–40.

Cui, H. & Heidorn, P.B. 2007. The reusability of induced knowledge for the automatic semantic markup of taxonomic descriptions. *J. Amer. Soc. Inform. Sci. Technol.* 58: 133–149. http://dx.doi.org/10.1002/asi.20463

Cui, H., Boufford, D. & Selden, P. 2010. Semantic annotation of biosystematics literature without training examples. *J. Amer. Soc. Inform. Sci. Technol.* 61: 522–542. http://dx.doi.org/10.1002/asi.21246

Fallside, D.C. & Walmsley, P. 28 Oct 2004. XML schema part 0: Primer second edition. W3C recommendation. http://www.w3.org/TR/2004/REC-xmlschema-0-20041028/ (accessed 26 Sep 2012).

Feigenbaum, L., Herman, I., Hongsermeier, T., Neumann, E. & Stephens, S. 2007. The semantic web in action. *Sci. Amer.* 297: 90–97. http://www.scientificamerican.com/article.cfm?id=semantic-web-in-actio. http://dx.doi.org/10.1038/scientificamerican1207-90

Freeland, C. 22 February 2011. Digitization and enhancement of biodiversity literature through OCR, scientific name mapping and crowdsourcing. BioSystematics Berlin 2011. http://www.slideshare.net/chrisfreeland/digitization-and-enhancement-of-biodiversity-literature-through-ocr-scientific-names-mapping-and-crowdsourcing (accessed 24 Oct 2013).

Hagedorn, G. 2007. *Structuring descriptive data of organisms: Requirement analysis and information models*. Dissertation, Universität Bayreuth, Germany.

IAWA Committee. 1989. IAWA list of microscopic features for hardwood identification. *I. A. W. A. Bull.*, n.s., 10: 221–332.

Kirkup, D., Malcolm, P., Christian, G. & Paton, A. 2005. Towards a digital African Flora. *Taxon* 54: 457–466. http://dx.doi.org/10.2307/25065373

Lapage, S.P., Sneath, P.H.A., Lessel, E.F., Skerman, V.B.D., Seeliger, H.P.R. & Clark, W.A. (eds.) 1992. *International Code of Nomenclature of Bacteria: Bacteriological Code,* 1990 Revision. Washington, D.C.: ASM Press.

Lydon, S.J., McGee Wood, M., Huxley, R. & Sutton, D. 2003. Data patterns in multiple botanical descriptions: Implications for automatic processing of legacy data. *Syst. Biodivers.* 1: 151–157. http://dx.doi.org/10.1017/S1477200003001129

Marhold, K., Stuessy, T., Agababian, M., Agosti, D., Alford, M.H., Crespo, A., Crisci, J.V., Dorr, L.J., Ferencová, Z., Frodin, D., Geltman, D.V., Kilian, N., Linder, H.P., Lohmann, L.G., Oberprieler, C., Penev, L., Smith, G.F., Thomas, W., Tulig, M., Turland, N. & Zhang, X.-C. 2013. The future of botanical monography: Report from an international workshop, 12–16 March 2012, Smolenice, Slovak Republic. *Taxon* 62: 4–20.

McNeill, J., Barrie, F.R., Buck, W.R., Demoulin, V., Greuter, W., Hawksworth, D.L., Herendeen, P.S., Knapp, S., Marhold, K., Prado, J., Prud'homme van Reine, W.F., Smith, G.F., Wiersema, J.H. & Turland, N.J. 2012. *International Code of Nomenclature for algae, fungi and plants (Melbourne Code): Adopted by the Eighteenth International Botanical Congress Melbourne, Australia, July 2011*. Regnum Vegetabile 154. Koenigstein: Koeltz Scientific Books.

Morrison, D.A. 2012. [Book review] *Tools for identifying biodiversity: Progress and problems*. *Syst. Biol.* 61: 710–712. http://dx.doi.org/10.1093/sysbio/sys007

Penev, L., Lyal, C.H.C., Weitzman, A., Morse, D.R., King, D., Sautter, G., Georgiev, T., Morris, R.A., Catapano, T. & Agosti, D. 2011. XML schemas and mark up practices of taxonomic literature. *ZooKeys* 150: 89–116. http://dx.doi.org/10.3897/zookeys.150.2213

Quin, L.R.E. 2010a. W3C: Standards: XML technology: Schema. http://www.w3.org/standards/xml/schema (accessed 26 Sep 2012).

Quin, L.R.E. 2010b. W3C: Standards: XML technology: XML essentials. http://www.w3.org/standards/xml/core (accessed 24 Oct 2012).

Ride, W.D.L., Cogger, H.G., Dupuis, C., Kraus, O., Minelli, A., Thompson, F.C. & Tubbs, P.K. 2000. *International Code of Zoological Nomenclature*. London: The International Trust for Zoological Nomenclature 1999 c/o The Natural History Museum.

Roos, M.C., Berendsohn, W.G., Dessein, S., Hamann, T., Hoffmann, N., Hovenkamp, P., Janssen, T., Kirkup, D., De Kok, R., Sierra, S.E.C., Smets, E., Webb, C. & Van Welzen, P.C. 2011. e-Flora Malesiana: State of the art and perspectives. *Gard. Bull. Singapore* 63: 189–195.

Sautter, G., Böhm, K. & Agosti, D. 2007. A quantitative comparison of XML schemas for taxonomic publications. *Biodivers. Informatics* 4: 1–13.

Seltmann, K.C., Pénzes, Z., Yoder, M.J., Bertone, M.A. & Deans, A.R. 2013. Utilizing descriptive statements from the Biodiversity Heritage Library to expand the Hymenoptera anatomy ontology. *PLoS ONE* 8(2): 1–7. http://dx.doi.org/10.1371/journal.pone.0055674

Sosef, M.S.M., Wieringa, J.J., Jongkind, C.C.H., Achoundong, G., Azizet Issembé, Y., Bedigian, D., Van den Berg, R.G., Breteler, F.J., Cheek, M., Degreef, J., Faden, R.B., Goldblatt, P., Van der Maesen, L.J.G., Ngok Banak, L., Niangadouma, R., Nzabi, T., Nziengui, B., Rogers, Z.S., Stévart, T., Van Valkenburg, J.L.C.H., Walters, G. & De Wilde, J.J.F.E. 2006. *Check-list des plantes vasculaires du Gabon = Checklist of Gabonese vascular plants*. Scripta Botanica Belgica 35. Meise: National Botanic Garden of Belgium.

Thessen, A.E. & Patterson, D.J. 2011. Data issues in the life sciences. *ZooKeys* 150: 15–51. http://dx.doi.org/10.3897/zookeys.150.1766

Thessen, A.E., Cui, H. & Mozzherin, D. 2012a. Applications of natural language processing in biodiversity science. *Advances Bioinformatics*, 2012: Article ID 391574. http://dx.doi.org/10.1155/2012/391574

Thessen, A.E., Patterson, D.J. & Murray, S.A. 2012b. The taxonomic significance of species that have only been observed once: The genus *Gymnodinium* (Dinoflagellata) as an example. *PLoS ONE* 7: 1–34. http://dx.doi.org/10.1371/journal.pone.0044015

Thomer, A., Vaidya, G., Guralnick, R., Bloom, D. & Russell, L. 2012. From documents to datasets: A MediaWiki-based method of annotating and extracting species observations in century-old field notebooks. *ZooKeys* 209: 235–253. http://dx.doi.org/10.3897/zookeys.209.3247

Thompson, H.S., Beech, D., Maloney, M. & Mendelsohn, N. 28 Oct 2004. XML schema part 1: Structures second edition. W3C recommendation. http://www.w3.org/TR/2004/REC-xmlschema-1-20041028/structures.html (accessed 26 Sep 2012).

Van Steenis, C.G.G.J. 1947. II. Instructions for cooperators. 5. Delimitation of the Malaysian region. *Fl. Males. Bull.* 1: 5–7.

Venin, M., Kirchhoff, A., Fradin, H., Güntsch, A., Hoffmann, N., Kohlbecker, A., Kuntzelmann, E., Maiocco, Ô., Müller, A., Vignes Lebbe, R. & Berendsohn, W.G. 2010. Descriptive data in the EDIT platform for cybertaxonomy. Pp. 7–11 in: Nimis, P.L. & Vignes-Lebbe, R. (eds.), *Tools for identifying biodiversity: Progress and problems*. Trieste: Edizioni Università di Trieste.

Weitzman, A.L. & Lyal, C.H.C. 2004. An XML schema for taxonomic literature – taXMLit. http://www.sil.si.edu/digitalcollections/bca/documentation/taxmlitv1-3intro.pdf (accessed 9 Apr 2013).