Content:

- 1. Introduction
- 2. Processing taxon descriptions in bulk.
 - a) Worked example
- 3. Adding individual taxon descriptions

1. Introduction

Published taxon descriptions from regional flora accounts, new species descriptions, partial or complete monographs represent a valuable resource in biodiversity informatics. Digitising these for incorporation in a scratchpad is can be a useful initial step in web-taxonomy. Permission to reproduce any published work needs to be sought from the publishers as appropriate. The advantages of carrying out this work are:

- it assembles relevant treatments of a taxon in one place
- the information is available online as opposed to inaccessible in the specialist literature
- the information is searchable because it is stored in a structured database

Scratchpads can hold multiple descriptions of the same taxon and the descriptions can be in different languages too.

Existing electronic resources:

http://www.biodiversitylibrary.org/advsearch http://plants.jstor.org/ (look under Collection for digitised floras) www.eFloras.org www.kew.org/efloras

How you process species descriptions depends on the format of the descriptions, access to equipment and software and the number of taxa treated in the manuscript. If there are several descriptions in a manuscript that all follow the same format, it is probably worth processing them in bulk. If there are fewer species descriptions it may be more efficient to copy and paste directly into the scratchpad.

If you have an electronic copy of the manuscript, you can ignore the first few steps of the processing described below. Permission to reproduce may still need to be sought. Please read through the whole of the document to get a better understanding of the principles involved at each step.

The version of Excel used in screenshots is Excel2007. The principles should be adaptable to earlier. See also scratchpad wiki: http://help.scratchpads.eu/w/Add and export taxon descriptions, http://help.scratchpads.eu/w/Import.

Please get in touch with the Content team or the Scratchpad team if you get stuck.

2. Processing Taxon descriptions in bulk

Basic steps: hardcopy the manuscript (book or journal article) \rightarrow PDF \rightarrow Word document \rightarrow UTF-8 encoded Text file → Excel file → TEMPLATE-import_into_taxon_description.

Workflow for Processing Taxon descriptions in bulk

Please note that Steps 2 and 5 are explained in more detail in the Worked example below.

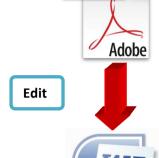


PDF

If the manuscript is not available electronically, you need to scan the hardcopy. We are scanning and saving pages individually as PDFs or JPEG images and then combining them into a single PDF file using Adobe Acrobat writer.

1. The scanned article needs to be converted into a machine-readable document – this process is called OCR (Optical Character Recognition). Adobe Acrobat writer has OCR fun Aa nality (). Another OCR program (and with a free trial version) is Abbyy FineReader which comes with a user manual.

Most modern PDFs are already machine-readable.



Encode

Import

2. Copy and paste the text from the PDF into a word document for proofreading and editing. The amount of proofreading depends on how clean the OCR was.

Edit the Word document and add delimiters. This involves firstly, proof-reading and editing the text that goes into the scratchpad. Secondly, it involves adding the delimiters (e.g. #) that separate the text for different fields (eg. Taxon name, General description, Habitat, etc.) in the scratchpad. See Worked Example below.

3. Save the final word document as plain text (*.txt), other encoding: Unicode(UTF-8).

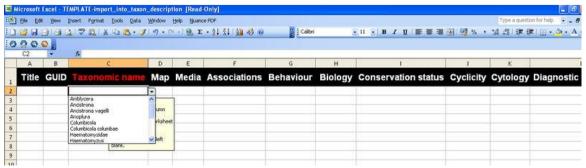
UTF-8 encoded text file. This is a format that can be imported into an excel sheet.

4. In excel, select get external data from a text file and navigate to the text file you made; import choosing 'delimited' data type (*select next*); set the delimiter to 'Other:' and add your chosen delimiter (i.e. #) in the provided box (*select next*), *select finish*, select the first cell in the sheet for putting the data.

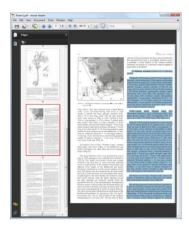
The taxon name and corresponding information on General description, Habitat etc in separate fields, display in the **Excel sheet**. This is then **mapped into the scratchpad TEMPLATE**. See Worked Example below.



5. Download a new copy of the TEMPLATE for taxon descriptions from your scratchpad. Then map the fields from the **Excel sheet** to the **TEMPLATE** as described in the Worked example 2a below .



2a) Worked example



←PDF with species descriptions. The text highlighted in blue is copied and pasted into a Word document for editing. Stages in the editing of the text are illustrated below in Box 1 (instructions 1-3) and Box 2 (final version of the text after editing).

Step 2. Edit the Word document and add delimiters.

Copy and paste the text form the PDF into a Word document and

- 1. correct spelling mistakes that have resulted from the OCR. In particular, check that the name of the taxon is spelled exactly the same as in the classification tree in your scratchpad (written out in full, with a single space between the genus name and the epithet).
- 2. delete any text you don't need.
- 3. add delimiters between parts of the text that need to be displayed in different fields in the scratchpad (e.g. General Description, Habitat, Phenology). # works well as a field delimiter because this character doesn't usually get used in the text. Add a different delimiter at the start of each new taxon ('# works well); this is required to make each taxon a separate record in the Excel sheet later on (see below).

field delimiter '# Taxon delimiter

If there is a notes section, you may need to split the text up so the information end up in the appropriate fields in the scratchpad. For example, a statement on how the species differs from another species can go in the diagnostic description field, whilst the discussion on the name can go into Taxonomic notes.

Distribution and habitat are often treated together. It's best to split the statements relating to distribution from those on habitat.

4. complete a number of find and replace actions as listed in table 1.

Table 1: Find and replace actions. The first three replacements need to be completed in the order they are listed below. The next two are optional replacements.

Find what:	Replace with:	Note
^p (paragraph	Space	This is replacing paragraph marks with a space.
mark)		
Double space	Single space	Repeat until there are no more replacements done.
' #	^p'#	This makes every taxon into a different paragraph. When the final text file is imported into excel this will be a different record.
X	×	e.g. 2 – 4 x 7 – 14. Tip: Include a space before and after the x and replace it with a space before and after the special character.
- (hyphen)	- (En dash)	e.g. 2-7. Ranges are commonly indicated by an en dash: 2 – 7.

66. "Babiana scabrifolia Brehm ex Klatt in Abhandlungen der Naturforschenden Gesellschaft zu Halle 15: 349 (Erganzungen 15) (1882); G.J.Lewis: 68 (1959) (Type: #South Africa, [Western Cape], Clanwilliam District, Langevallei, July 1830, JF Drege 2623 [K, 1;:cto., designated by Lewis, 1959: 68; P, iso. (two sheets)].

B. scabrifolia var. acuminata G.J.Lewis: 70 (1959). Type: South Afri ca, [Western Cape], Olifants River Valley at Bulshoek Barrage, 2 August 1950, G.J. Lewis 2207 (SAM, bolo.; SAM, iso.).
B. scabrifolia var. dec/ina/a G.J.Lewis: 71 (1959). Type: South Africa, [Western Cape], between Brandewyn River and Doorn River, in sa nd, 25 August 1950, G.J. Lewis 2192 (SAM, bolo.).
See Lewis (1959) for complete synonymy.

#Plants 50- 150 mm high including leaves, with a thick collar of fibres around stem base, \pm acaulescent, often forming tufts, simple or branched at ground level, hairy above ground level. Leaves lanceolate to oblong, exceeding stem, 60-100 x 6-20 mm, usually inclined toward ground, soft-textured and lightly pleated, shorthairy or nearly smooth, narrow and twisted in young plants. Bracts 20- 32 mm long, green, sparsely hairy, the inner slightly shorter than the outer, divided to base (rarely shortly above base), with wide transparent margins. Flowers zygomorphic, 4-8 in a dense, inclined to horizontal spike, mostly pale blue to pale lilac, lower lateral tepals with broad white to creamcoloured splashes in upper hal f outlined in dark blue to violet, sweetly scented, often of narcissus; perianth tube narrowly funnel-shaped, 12-18 mm long; tepa Is unequal, dorsal 30-45 x 7-10 mm, lower tepals 20-30 mm long. joined to upper laterals for up to 4 mm and to one another for ± 4 mm forming a prominent lip, margins of lower laterals undulate, sli ghtly crisped in lower half. Stamens

unilateral; filaments arched, 13-18 rnm long; anthers 6-8 mm long; pollen white. *Ovary* thinly hairy above or on ribs, rare ly smooth; style dividing between middle and apex of anthers, branches 4-5 mm long, expanded at

tips. Flowering time #June to August. Plate 58.

Distribution and ecology #Western Cape: in the Olifants River Valley and lower slopes of the surrounding mountains; #stony soils in dry fynbos or karroid scrub (Map 36).

Hardly differing in its flowers from several other species of section Babiana, B. scabrifolia is most easily recogni zed by the underground stem, often branched at ground level, and usually forming small tufts. It reproduces amply by cormlets and it is common to see mature flowering plants with their broad, soft-textured leaves surrounded by the linear, twisted to coiled leaves produced by immature corms. The flowers are relatively large, and although oveltopped by the leaves, make an attractive sight in the Olifants River Valley and surrounding bills in early spring. The flowers have a pleasing sweet scent and contrast both in the scent and large size from those of B. mucronata, which is also common in the Olifants River Valley. That species typically has an erect, usually well-developed aerial stem, whereas the flowers have a densely hairy ovary and produce a rather harsh acrid-spicy odour (sometimes described as flea powder), and does not have young plants with narrow, twisted to coiled immature leaves surrounding the parent plant.

← Box 1: Text copied from a PDF file, showing the workings for the first three instructions listed above. Deletions are in 'strikethrough'. Circled in green are the headers of different types of additional information about the species. These need to be deleted too since the text is assigned to the corresponding field in the scratchpad during mapping (step 5) at the end of the work flow. Words that need to be corrected because of incorrect OCR are highlighted in yellow. See finished text in Box 2.

[&]quot;#next species #TYPE INFO#descriptions new sp 1 #flowering time sp. 1 #distribution sp 1 # #differs from sp. 2.

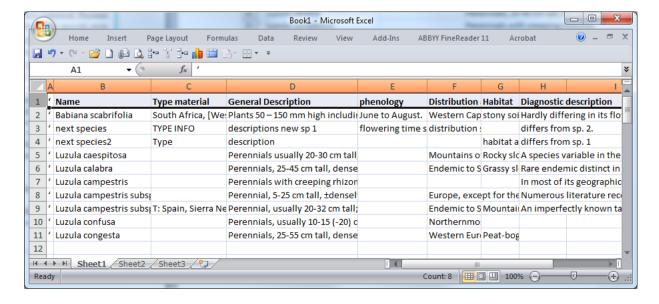
[&]quot;#next species2 #Type #description ## #habitat and no text for flowering time and distribution #differs from sp. 1

Box 2: Text from the PDF file above with all the editing complete. Note the corrections and the character replacements compared to Box 1. Exercise: copy and paste the text in the box into a new word document and save it as a plain text file encoded as Unicode (UTF-8). Then import it into an excel file (see instructions under step 4 above) and notice how the text has spread across different fields. Note the necessity for multiple delimiters if there is no text for some of the fields.

"#Babiana scabrifolia #South Africa, [Western Cape], Clanwilliam District, Langevallei, July 1830, JF Drege 2623 [K, lecto., designated by Lewis, 1959: 68; P, iso. (two sheets)]. #Plants 50 – 150 mm high including leaves, with a thick collar of fibres around stem base, ± acaulescent, often forming tufts, simple or branched at ground level, hairy above ground level. Leaves lance of ate to oblong, exceeding stem, $60 - 100 \times 6 - 20$ mm, usually inclined toward ground, soft-textured and lightly pleated, short-hairy or nearly smooth, narrow and twisted in young plants. Bracts 20 – 32 mm long, green, sparsely hairy, the inner slightly shorter than the outer, divided to base (rarely shortly above base). with wide transparent margins. Flowers zygomorphic, 4-8 in a dense, inclined to horizontal spike, mostly pale blue to pale lilac, lower lateral tepals with broad white to cream-coloured splashes in upper hal f outlined in dark blue to violet, sweetly scented, often of narcissus; perianth tube narrowly funnel-shaped, 12 - 18 mm long; tepals unequal, dorsal 30 - $45 \times 7 - 10$ mm, lower tepals 20 - 30 mm long, joined to upper laterals for up to 4 mm and to one another for ± 4 mm forming a prominent lip, margins of lower laterals undulate, slightly crisped in lower half. Stamens unilateral; filaments arched, 13 – 18 mm long; anthers 6-8 mm long; pollen white. Ovary thinly hairy above or on ribs, rare ly smooth; style dividing between middle and apex of anthers, branches 4 – 5 mm long, expanded at tips. #June to August. #Western Cape: in the Olifants River Valley and lower slopes of the surrounding mountains #stony soils in dry fynbos or karroid scrub. #Hardly differing in its flowers from several other species of section Babiana, B. scabrifolia is most easily recognized by the underground stem, often branched at ground level, and usually forming small tufts. It reproduces amply by cormlets and it is common to see mature flowering plants with their broad, soft-textured leaves surrounded by the linear, twisted to coiled leaves produced by immature corms. The flowers are relatively large, and although overtopped by the leaves, make an attractive sight in the Olifants River Valley and surrounding bills in early spring. The flowers have a pleasing sweet scent and contrast both in the scent and large size from those of B. mucronata, which is also common in the Olifants River Valley. That species typically has an erect, usually well-developed aerial stem, whereas the flowers have a densely hairy ovary and produce a rather harsh acrid-spicy odour (sometimes described as flea powder), and does not have young plants with narrow, twisted to coiled immature leaves surrounding the parent plant.

Step 5: Excel sheet for mapping into TEMPLATE

The Excel sheet you get should look similar to this (add a row at the top to say what field the data is):

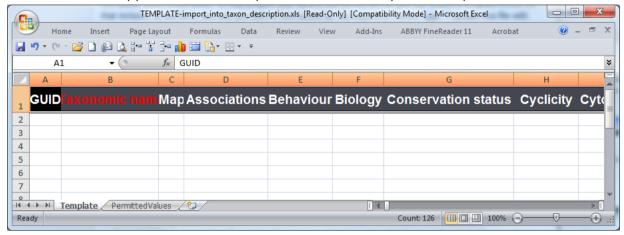


^{&#}x27;#next species #TYPE INFO#descriptions new sp 1 #flowering time sp. 1 #distribution sp 1 # #differs from sp. 2.

^{&#}x27;#next species2 #Type #description # # #habitat and no text for flowering time and distribution #differs from sp. 1

You need to look through the data in the Excel sheet carefully and make sure that all the data is in the appropriate column. Failing to do so will mean that data ends up in the wrong field in the scratchpad. Tips: sort the data in the different fields in alphabetical order. This may help to quickly spot any anomalies, especially if there is data missing that you would normally expect to be there. If that is the case, look in the adjacent fields to the left (all of them) to see if you can find the data there – there may be a delimiter missing in the text file. Make any necessary corrections in the text file and save it. Then import the text file into Excel again. That will be the final **Excel sheet** to use for mapping.

Download a fresh copy of the taxon description TEMPLATE from your scratchpad:

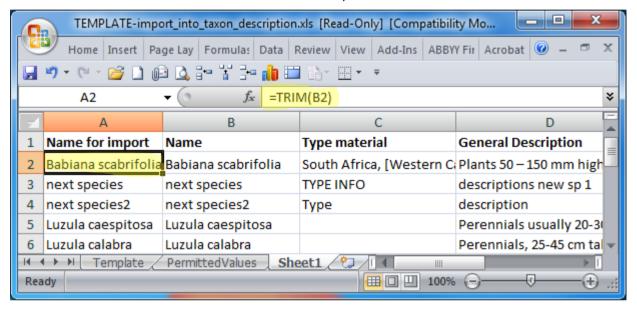


Four things to notice:

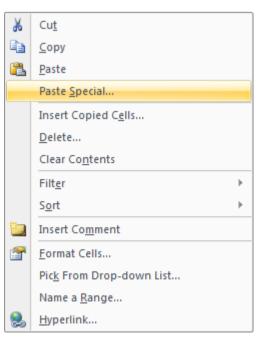
- 1. the two sheets in the document one called 'Template' and the other one called 'Permitted Values.
- 2. the number of fields (or headings) in the Template worksheet, most of which you don't need.
- 3. the Taxonomic names field and the reference fields are set up to only allow the values listed in the Permitted Values worksheet. There are only permitted values for the references if there are fewer than 700 records in the bibliography. If there are
- 4. The reference fields as you scroll to the right in the Template sheet. Each field that can have content from the published literature needs to be referenced. There are two types of reference fields: (NID) and (Title). They correspond to a number called node ID or to the title of the bibliographic record that references the content. You will need to enter a value in one of the reference fields for each field with content (see step e below).

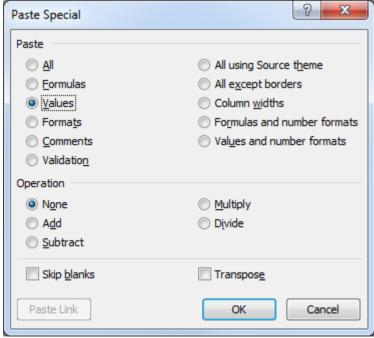
Then work through the steps below:

- a. Add your Excel document to the TEMPLATE as a new worksheet: click on 'Insert Worksheet' next to Permitted Values sheet (this should add a sheet called Sheet1); copy the worksheet in your **Excel document** and paste it into the new sheet in the TEMPLATE.
- b. Add a new column next to the species names. Call the new column 'Names for import' and populate it with the names in the existing Names column using the function: =TRIM(text). Use Auto fill to copy the function to all cells in the column. This function gets rid of superfluous spaces before or after the name. This is indispensible because a space at the start or the end of a taxon name would cause the scratchpad to identify it as a new addition to the classification. (see screenshot below where A2 returns the name in B2).

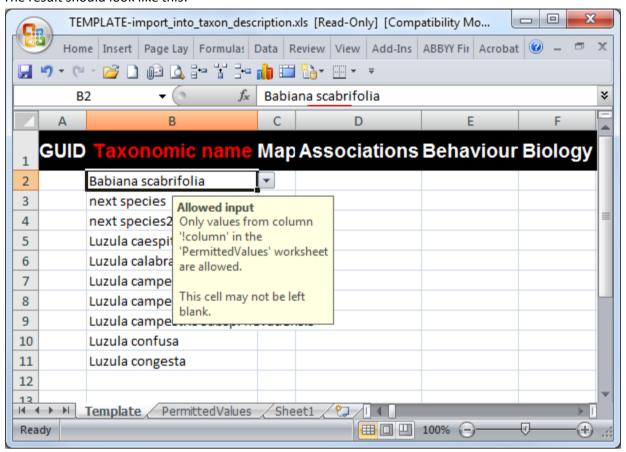


c. In Sheet1, select and copy all the names in the 'Names for import' column you made earlier. Switch to the Template sheet and right click into the field B2 (Just below the field entitled Taxonomic Name). Select Paste Special. This opens up a menu. Select 'Values' under Paste and click ok. This pasts the names of the taxa rather than the =TRIM(text) function.

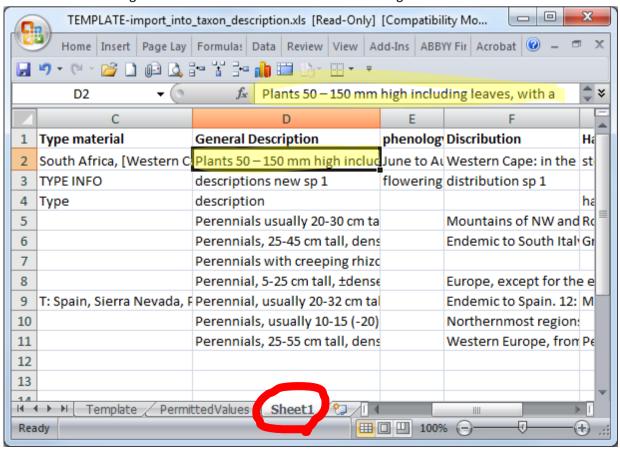


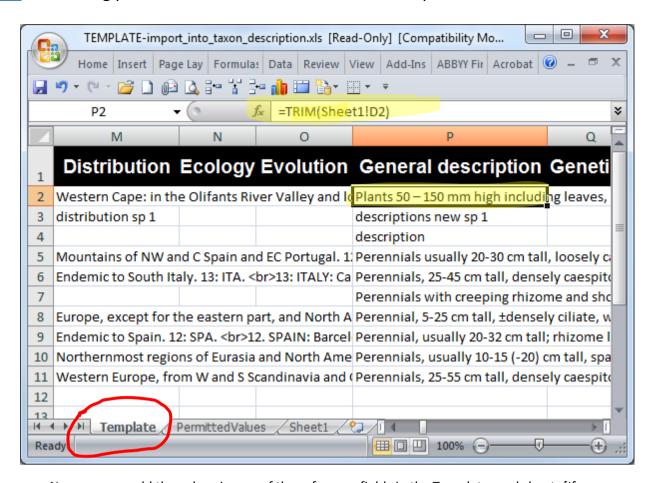


The result should look like this:

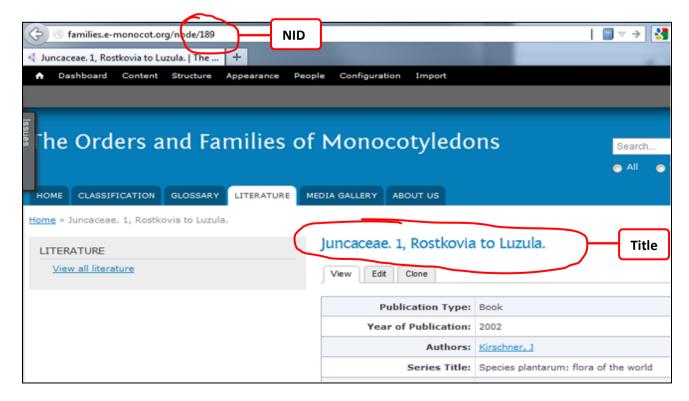


d. Next, map the remaining fields from Sheet1 to the Template worksheet using the =TRIM(text) function. Compare the next two screenshots: in P2 of the Template worksheet, the formula refers back to D2 in Sheet1 and returns the text without any superfluous spaces. The reason for this has to do with adding the references for records that have missing data.





e. Now you can add the values in one of the reference fields in the Template worksheet. [If you don't already have a bibliographic record for the publication from which you got the species description, you need to create one in the scratchpad.] Look up the Title or the node ID of the bibliographic record in the scratchpad or start typing the Title of the reference in the Title reference field and it should look up the reference in the Permitted Values drop-down. I will use the Title in the example below.



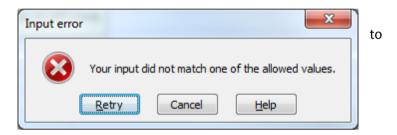
Only the fields that contain data must be referenced; so empty fields should not have a reference. In our experience, the General description field is the field you are most likely to have content for all the taxa in the worksheet and all the other content most likely comes from the same source.

If this is the case, we suggest referencing the General description field first. Copy and paste the Title from the Bibliographic record into the General description reference (Title) field*. Then apply the =IF([logical text], [value if true], [value if false]) function to populate the reference fields for the other content*. This avoids adding a reference for an empty content field.

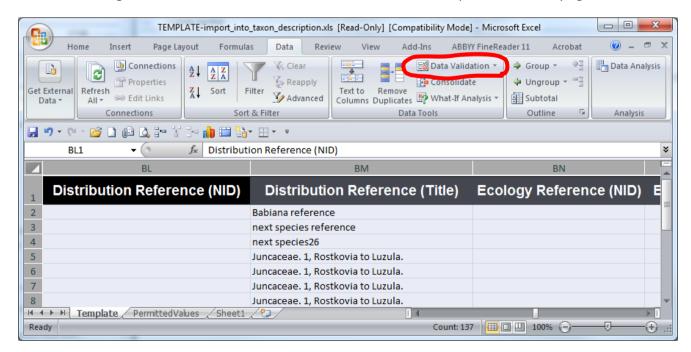
In our example, our function for the Distribution Reference (Title) field would be:

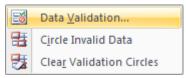
where M2 is the Distribution field and BS2 is the Reference Title field for General Description. The formula reads: if the Distribution field M2 is empty, please leave this field empty. If it is not empty, please fill it with what's in the General description Reference (Title) field, BS2.

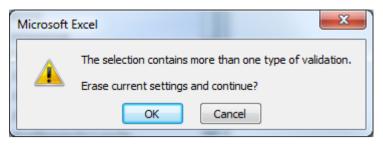
*If you get an error message like this it's because the reference you've tried use is not listed in the Permitted Values worksheet or because the validation rules don't allow functions. I suggest you change the validation rules as detailed below.

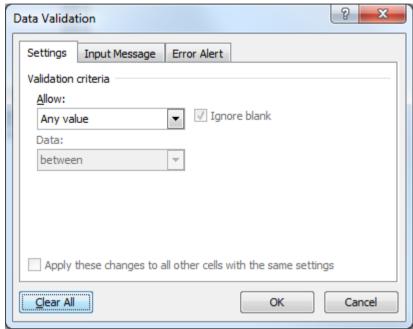


Highlight all the reference columns, select 'Data Validation' under the Data tab, say yes to 'Erase current settings and continue?', in the next window select clear all and press OK. Now try again.









- f. Now you need to replace any formulas in the worksheet with the actual text and finally, the file that you save and that you use in the import needs to only have the two worksheets 'Template' and 'Permitted values'.
 - Download a new TEMPLATE again (in particular if you've had to clear all data validation settings), select the all the records in the first Template worksheet, copy, paste into the new Template worksheet using the paste special: Values option. Save file. This file is now ready for import.

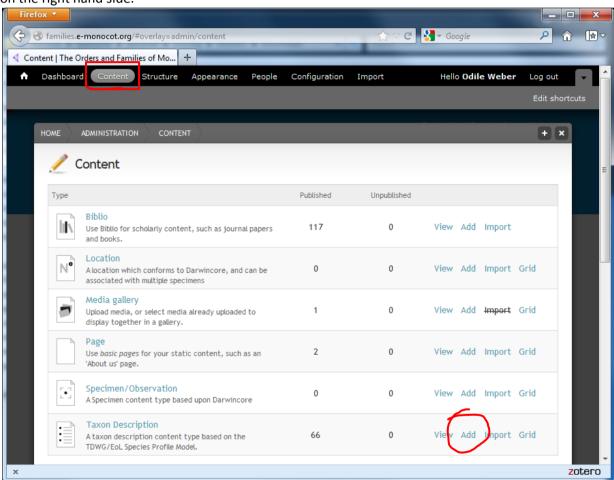
Remember to get in touch with the Content team or the Scratchpad team if you get stuck.

3. Adding individual taxon descriptions

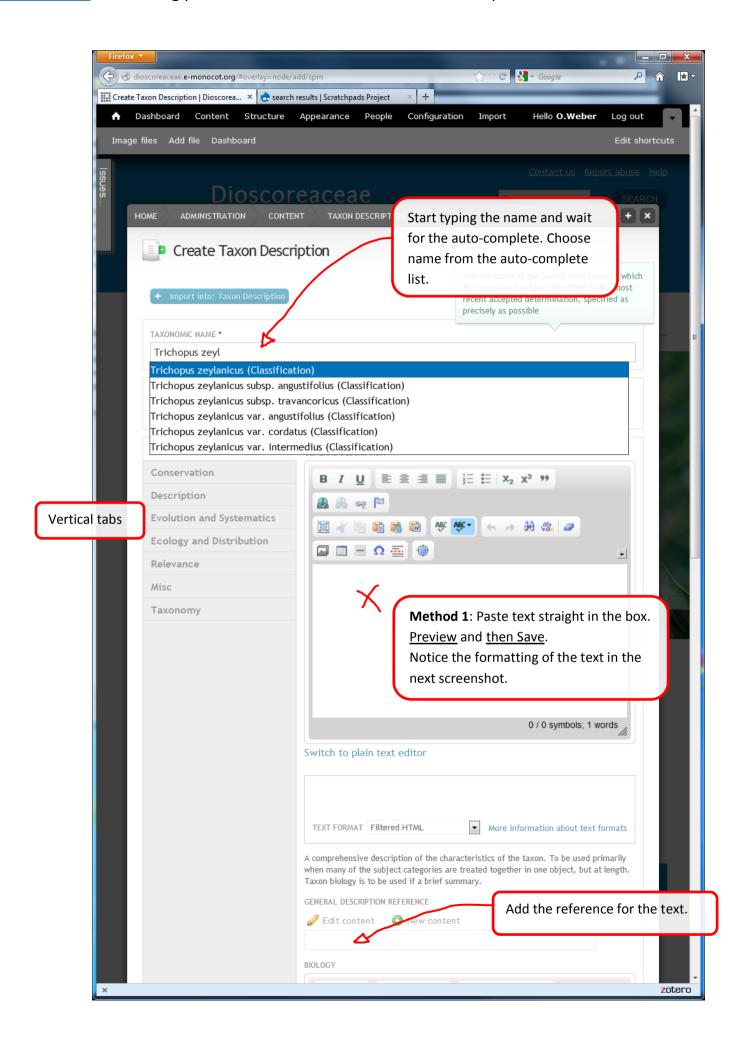
There are several different ways for copying and pasting taxon descriptions and associated information into the scratchpads. You may need to play around and decide for yourself which one works best for you, depending on whether you want to retain formatting in the text (e.g. Paste from Word).

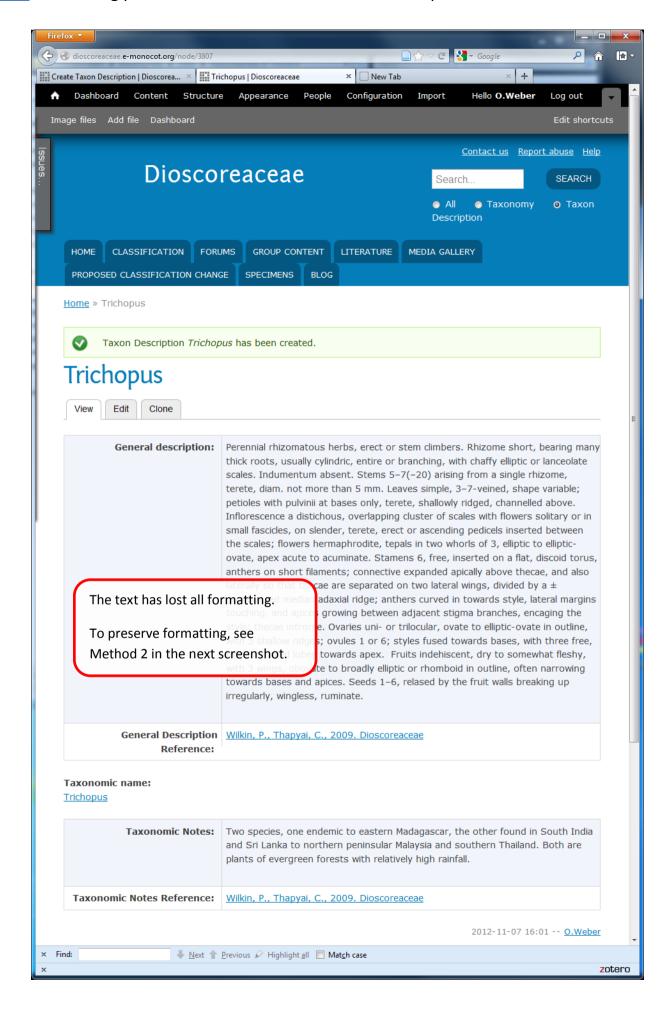
Within the text you copy over, you need to delete references to figures and maps. Each text field has a corresponding reference field, usually right below the text field. The reference fields are autocompleting (like the Taxonomic Name field) but sometimes it's difficult to find the reference from the drop-down list. If that's the case, search for the bibliographic record in a new tab and get the NID (see screenshot under Worked example above, step e). Paste the NID in the following format: [nid:###] (i.e. [nid:189] in the example above).

In the top bar, select Content tab. In the window that opens find Taxon Description and select Add on the right hand side.

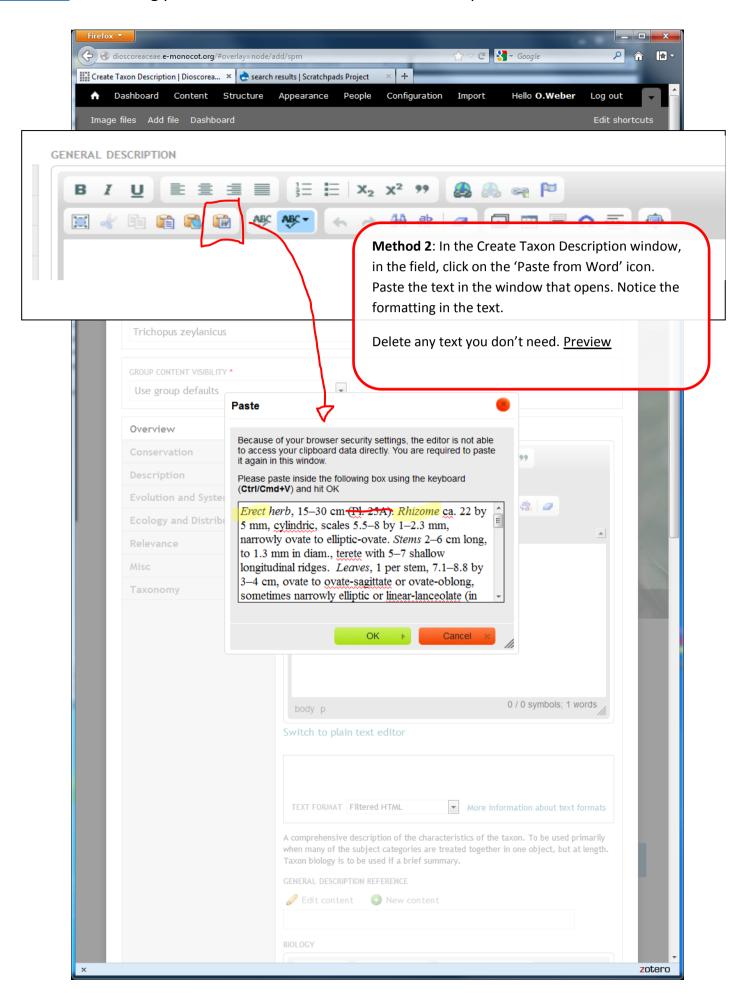


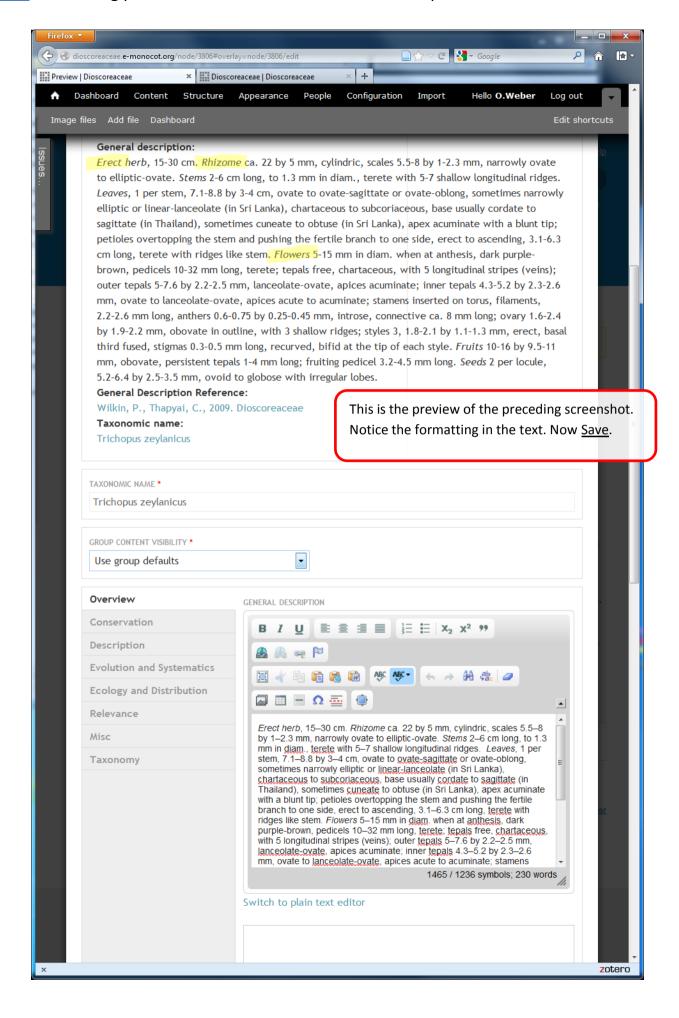
In the Create Taxon Description window (screen shot below), the fields are organised in different vertical tabs. Overview tab includes the Taxonomic Name field and the General Description field. Type out part of the name of the taxon in the Taxonomic Name's field. A drop-down list should appear with name options available from the Classification. Select the name you want. Paste the species description directly into the General description field.

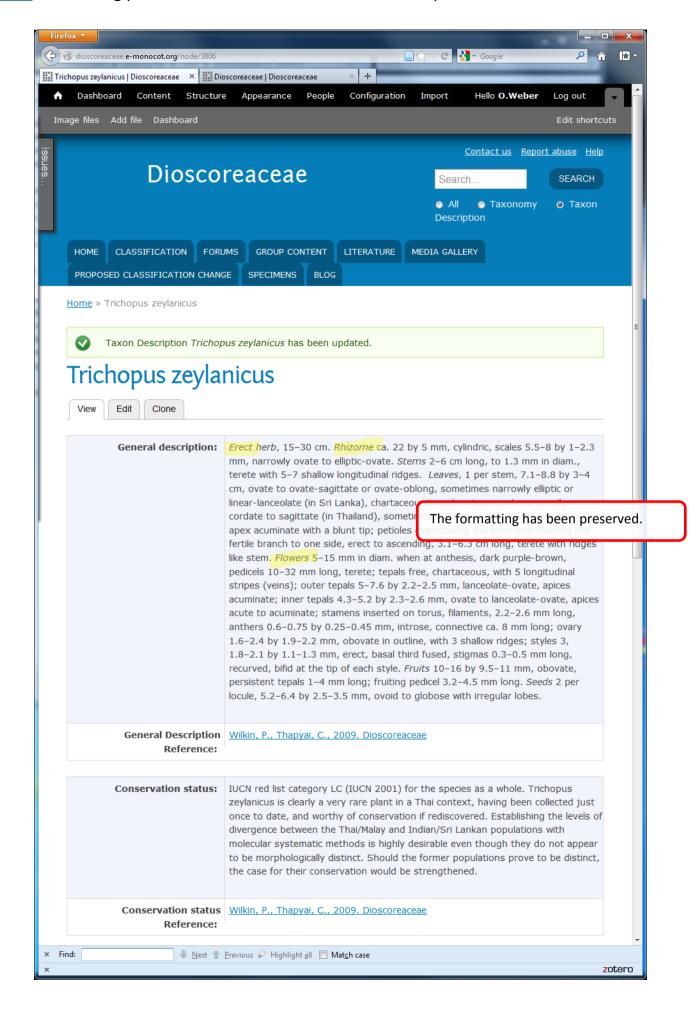


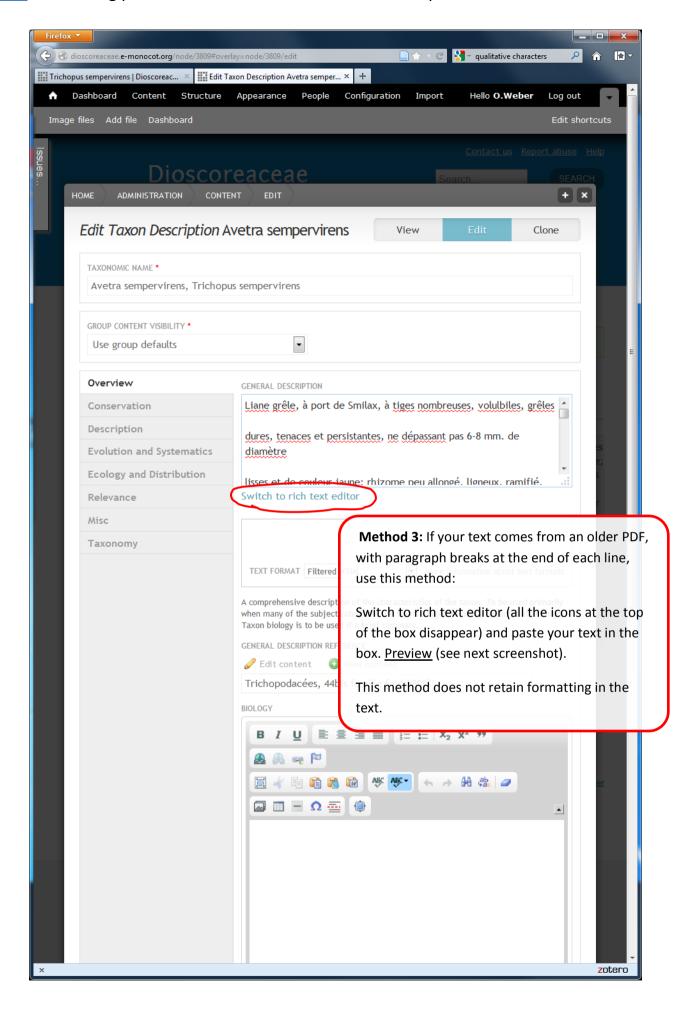


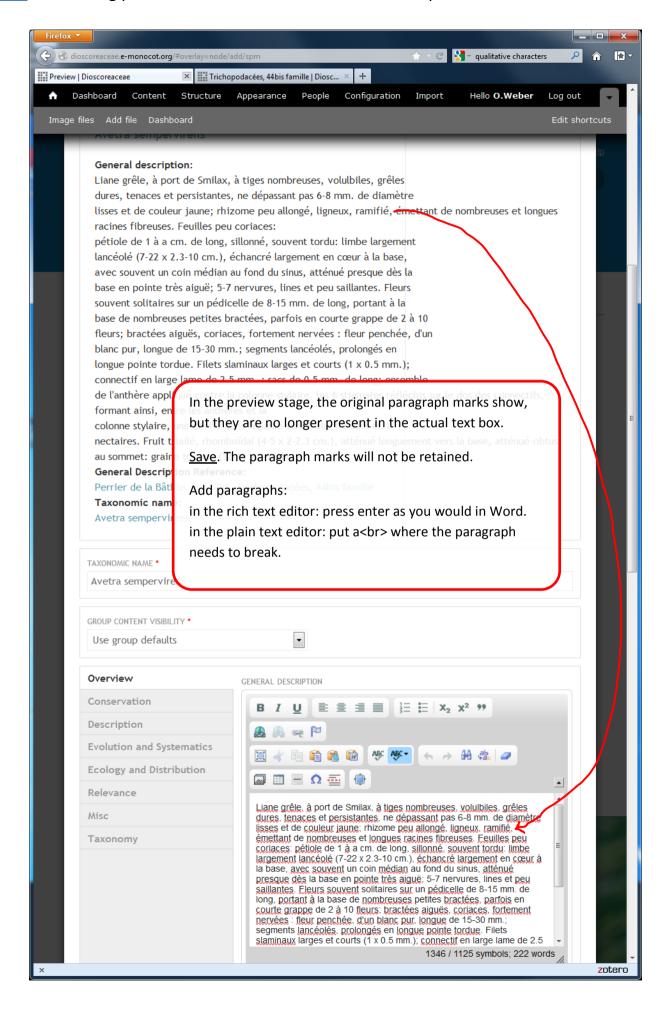


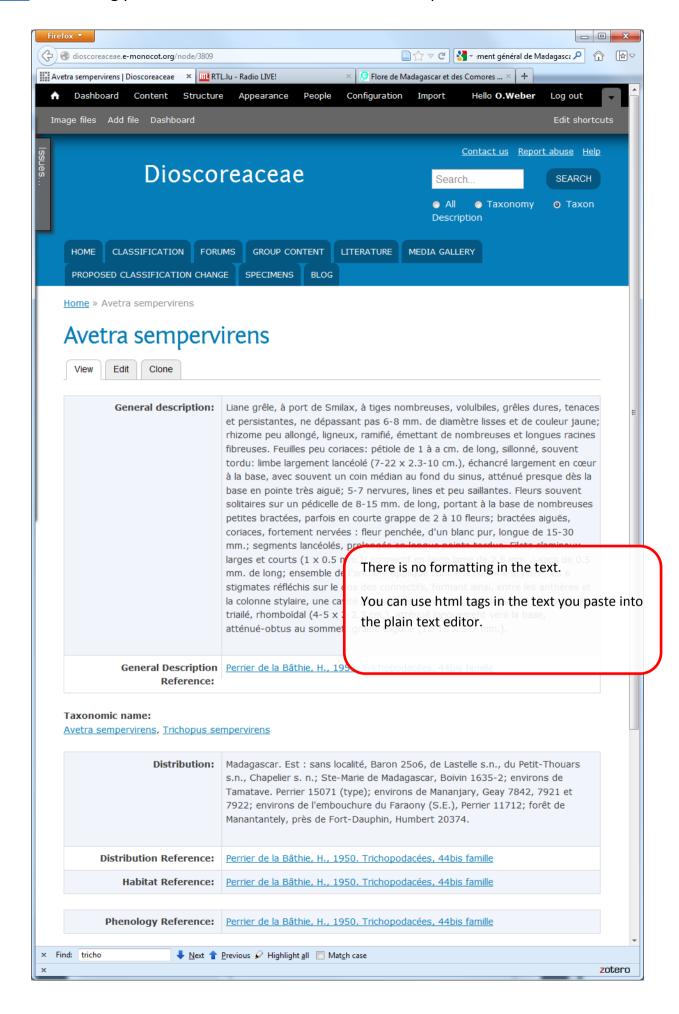












Remember to get in touch with the Content team or the Scratchpad team if you get stuck.