Image processing and metadata addition for online e-Flora use.

ver. 1.2

Thomas Hamann

Abstract: Flora images have to be processed separately from the Flora text before they are used in an online e-Flora. In this document the current procedure used for Flora Malesiana, Flore du Gabon, and Flora of the Guianas is described in detail, providing screenshots and tips to improve the quality of the images where needed, including local image administration using Microsoft Excel. Furthermore, the steps required to add metadata to each of the images are shown, and it is describe how the images are linked to their respective locations in the marked up treatments.

This project was subsidized in part by the EU project “pro-iBiosphere” (Grant agreement 312848).

Acknowledgements: I would like to thank Walter Berendsohn and Dominik Röpert for providing the information on the image processing tools used, and Andreas Müller and Gregor Hagedorn for their help and suggestions to improve the image processing procedure.

Table of Contents

Preface

The procedure described in this manual may seem low-tech and, perhaps, even somewhat simplistic, from a certain point of view. This has a reason.

When I first started to do this work back in 2010-2011, we were not sure yet how images would be hosted on the web. In the CDM database? Or perhaps on a dedicated server? Or would we use Naturalis Biodiveristy Center's then-not-yet developed image database? We didn't know. Furthermore, the Nationaal Herbarium of the Netherlands itself (this was in the pre-merger period with Naturalis) also did not have a suitable service available. Nor did Leiden University at the time. We also did not have the knowledge or resources to set up such a system ourselves.

This posed a problem.

Flora Malesiana has close to 4,000 (!) images. Some kind of system had to be set up to administer the information about those images, certainly before hosting and, due to the large number of images and the length of the flora itself, preferably also before actually extracting the images from the scans. It was a question of scale. And of how long one would want to work on a single task. Extracting the images from the scans and cleaning them up is repetitive work that is visually tiring. 4,000 images in one go takes too long time, several weeks at least. One has to take into account the potential for errors.

Furthermore, the structure of the flora called for a system that allowed references to figures and plates in the text to be linked to the actual images, even if the images themselves were not ready yet. Not doing so would create another problem, namely not knowing which figure reference would reference what image. Sure, a human could do that, but figure references generally start at 1 again for each new family treatment. This would confuse an automated script. On top of that, back then I couldn't script.

Setting up an image administration would require individually tracking each image. Linking image references to the actual images would also require this.

Why not combine the two problems, and make them a single problem, I thought?

XML has this cute little thing where one element can include a reference to an identifier and another element can include the identifier itself. What if I make that identifier the same number as the one I use to identify each image in the image administration?

Good. That *might* work.

This idea provided the genesis of the procedures described in this manual. I chose Microsoft Excel, the "poor man's database", as my instrument of choice for the image administration. It was a program (pardon, an "app") that I had used a lot in the past,

and was at ease with. The other option would have been Microsoft Access, which I had never used for serious work. Excel seemed “good enough”. The program to stuff metadata into images was suggested to me by the good people at the Botanic Gardens Botanic Museum Berlin, and could be driven using a batch file, another thing I had experience with. The rest, really, was purely a matter of “What do we need, and what is useful?”.

So.

I hope that this preface gives you a bit more insight on why things are as they are, and, likely, it also gives you some insight in how I think. Of course, I hope someone will make use of this manual, one day. It certainly was one of the most educating things I’ve ever written, together with the other manuals. I hope you like it.

Thomas Hamann, March 2016

(originally, the manual was written in 2013)

Introduction

Digitizing a legacy taxonomic work for use on the semantic web consists of two major tasks. The first task is digitizing of the text, usually by using optical character recognition (OCR) and XML mark up. However, each taxonomic work also contains images; figures, plates, photographs and other contents that are not text. These have to be processed separately from the treatment text before they can be used in an online e-Flora.

Merely making digitally available the scans of the pages of the legacy taxonomic treatment that feature these images is not enough and also riddled with problems. Not only the scans often are large in file size, making them long to download, but their file format may not be suited for use on the internet. Furthermore, scans will include unneeded text, which looks unprofessional, and may have quality problems such as stains. Finally, the images have to be made decently tamper-proof and have to be somehow linked to the treatment text. To make the last part possible, information about the images contained in the treatment text has to be embedded into the image files themselves. By doing this, it becomes possible to retrieve the image information even when the file name is accidentally altered.

This document aims at informing you of the method used to prepare Flora images for use on the web. This method consists of three major steps. First, an Image Administration is set up to which image information, metadata, can be added. Second, the image files are prepared for web use by processing them in Adobe Photoshop. Third, a batch file is created from the information contained in the Image Administration and used to add the metadata to the image files using a tool that can modify the metadata embedded in those files. This results in image files with metadata (Figure 1). The metadata also includes information that allows the image files to be linked to the XML file of the treatment.

The various steps are explained in detail in the rest of this manual.

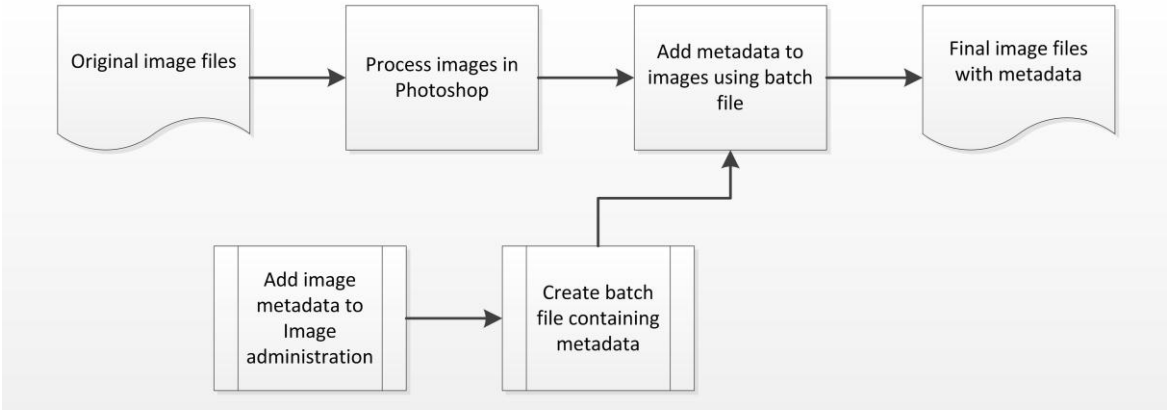


Figure 1: Global overview of procedure followed.

Preparations

Before any images can be prepared for use in an e-Flora, certain tasks have to be completed. In this section it is explained which computer programs and source files you must have available, which settings are suggested for some of the programs, and how a basic image file administration can be set up.

Required programs and tools

The following computer programs are used for the image preparation process:

- A file manager, such as Windows Explorer.
- Microsoft Excel or equivalent
- Adobe Photoshop CS2 or higher (not Photoshop Elements[1])
- Notepad++ (http://notepad-plus-plus.org/)
- EXIFTool (http://www.sno.phy.queensu.ca/~phil/exiftool/)

Furthermore, you should be able to run Batch files in Windows. This may require additional Windows usage rights; ask your system administrator or helpdesk.

It is assumed you are running a version of Microsoft Windows, preferably XP or higher.

If you are not running Windows, and instead are a user of Mac OS X, Linux, or any other UNIX-like operating system, you will need to find some replacement programs, and you will be on your own to make everything work. However, hopefully this manual is clear enough to get you started.

Windows set up

Windows should be set up properly to facilitate your work.

1) Go to the “Control Panel”, and check in the “Folder Options” control panel under the “View”-tab (Figure 2) that “Show hidden files, folders, and drives” is checked, and that “Hide extensions for known file types” is unchecked. If batch files remain invisible, you should also uncheck “Hide protected operating system files”.

2) In the “Region and Language” control panel, under the “Keyboards and Languages”-tab, choose “Change keyboards” and ensure that you are using a “US” keyboard layout, not the “US-International” keyboard layout (Figure 3).

[1] Photoshop Elements may not have all the functions required.

This disables the smart quote and accents feature that makes typing code cumbersome.

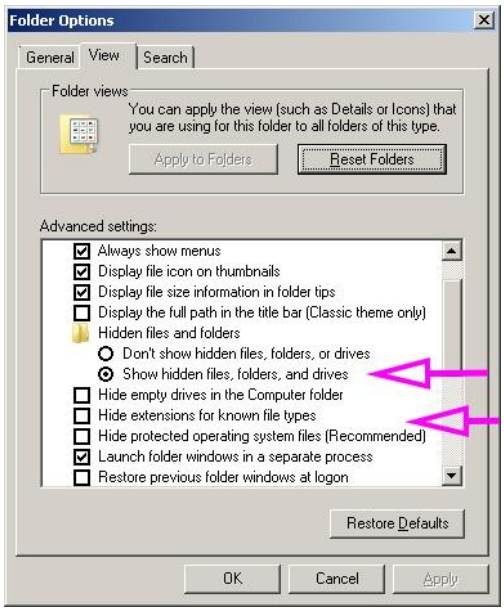


Figure 2: Important options to set in the "Folder Options" control panel.

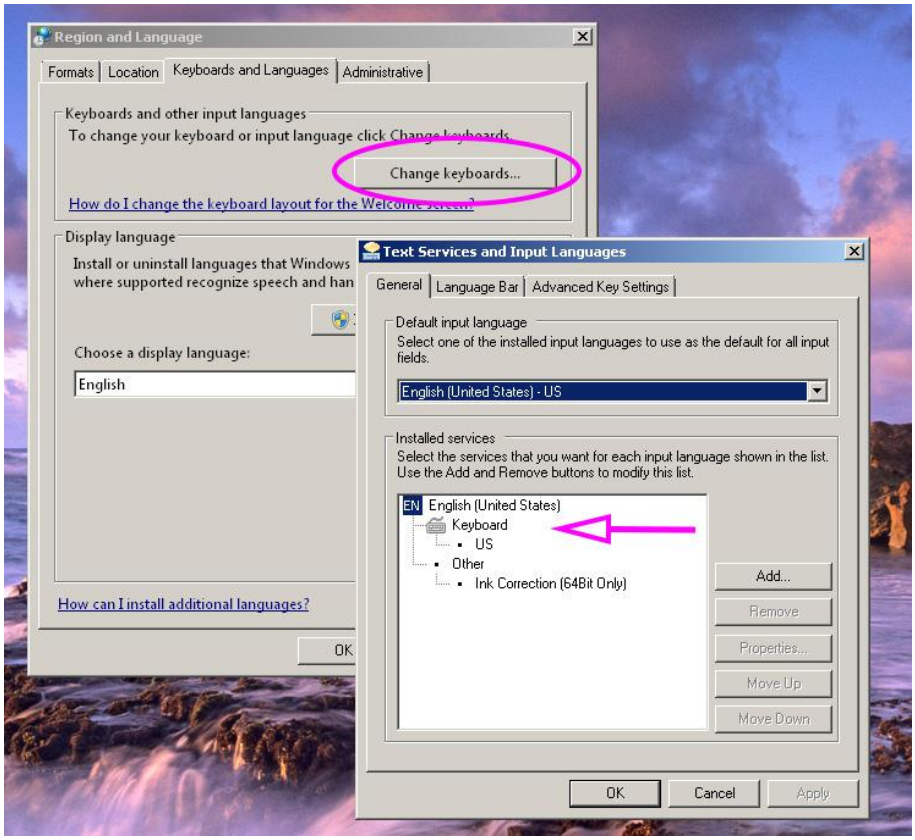


Figure 3: Proper keyboard settings.

Adobe Photoshop set up

Before you get started, some settings in Adobe Photoshop should be checked and even-tually changed. Adobe's settings can be found by going to the "Edit" menu and choosing "Preferences => General".

The options listed below are for Adobe Photoshop CS6. Other versions of Photoshop may have slightly different options; consult the "Help"-function if you encounter any trouble.

The following options are suggested:

- In the "File Handling" tab, uncheck the checkbox next to "Save as to original folder".
- In the "Cursors" tab, under "Painting cursors" choose the "Normal brush tip" option, check the checkbox next to "Show crosshair in brush tip", and choose the "Precise" option for "Other cursors".
- In the "Units & Rulers" tab, use centimetres, millimetres, or pixels as the unit for the "Rulers".

Furthermore, the workspace in Adobe Photoshop (essentially, what you see on your screen) should be set up. To do this properly, you might have to create a new file, because some options may otherwise not be available.

- In the "View" menu, enable the "Rulers" (there should be a check mark in front of the item).
- In the "View" menu, disable "Snap".
- In the "Window" menu, enable the "Actions", "Info", "Navigator", and "Histogram" pan-els. If required, these can be docked to the panels that are already present (Search Adobe Photoshop Help for "docking panels" to discover how to do this).

Figure 4 shows how your workspace could look by this point (colours may vary depending on version of Photoshop).

(continued next page)

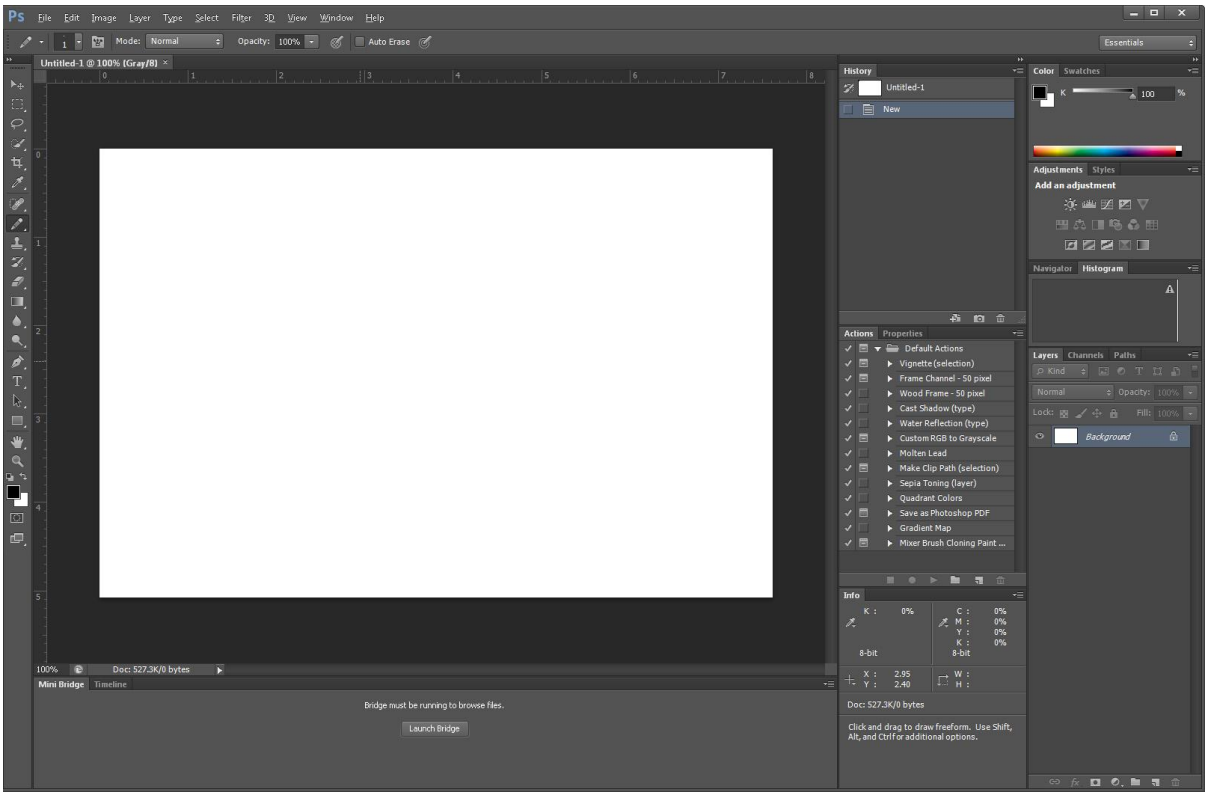
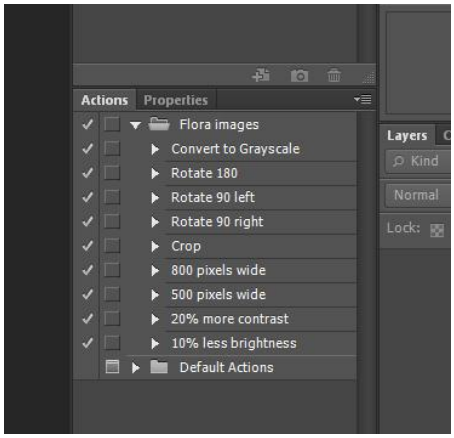


Figure 4: Adobe Photoshop workspace showing rulers and various panels.

Automation in Adobe Photoshop

When you will be processing image files in Photoshop, you will often be repeating the same tasks over and over again. This can be a source of errors and incorrect image file manipulation, which in some cases may lead to a degradation of the image quality.



To avoid this, you can use Photoshop's ability to record ***Actions***. Actions are macros, small computer programs containing one or more tasks that you can run at will. Adobe Photoshop's Help function explains quite well how Actions can be recorded. Be sure to give your Photoshop Actions meaningful names, i.e. ones that describe the action performed. Actions are saved in your Photoshop workspace, and can be reused at a later time.

Figure 5: Useful Photoshop Actions.

The following tasks are very suitable to be recorded as Actions: Image Mode Change, Cropping, Rotating, Resizing to various sizes, and modifications to the brightness and contrast of the image (Figure 5).

Tip: It is possible to run multiple Actions in sequence, which can seriously speed up the image processing time. To do this, select the first Action to perform, then press the “CTRL” key on your keyboard, and select the other Actions to perform (in sequence). In Figure 6, the Action “Convert to Grayscale”, “Crop”, and “800 pixels wide” have been selected after part of the image was selected. When you click on the “Play” button, all of these actions will be performed in one go, and the image will be ready to save.

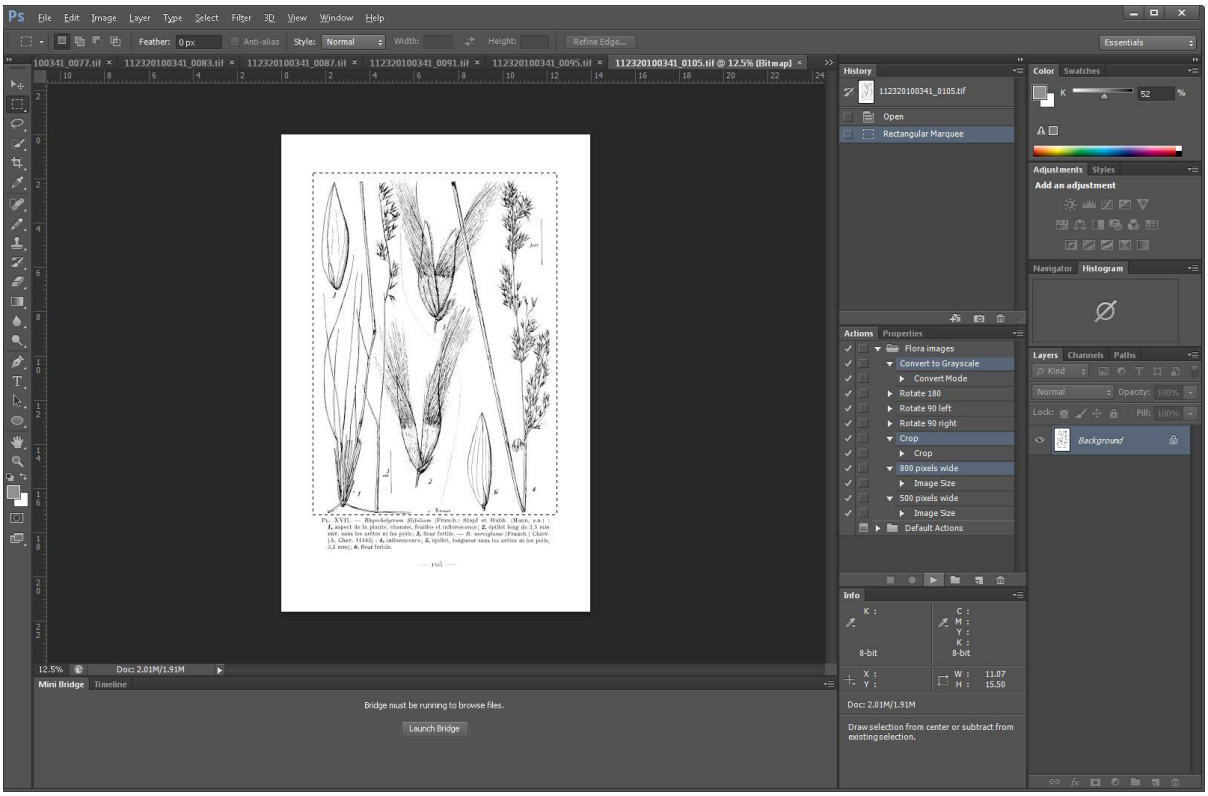


Figure 6: Running multiple Actions in sequence.

EXIFTool set up

Right-click the EXIFTool application in Windows Explorer and select “Properties”. Then change the file name in the “General” tab from “exiftool(-k).exe” to “exiftool.exe” so EXIFTool will run properly (Figure 7).

Furthermore, EXIFTool should be present in the folder from which you will run the batch process (see later). This means that the application should be in the folder, not a short-cut to the application.

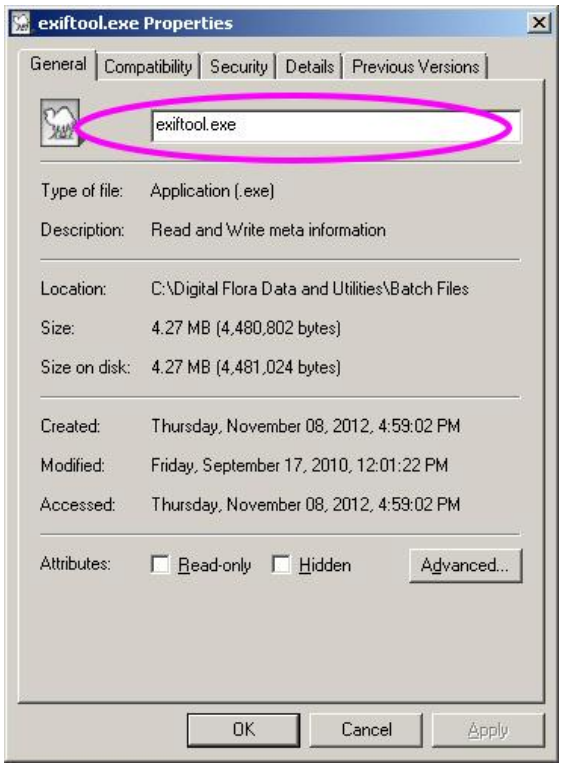


Figure 7: EXIFTool set up.

Source documents

The image files can have three different sources, given here in order of preference (highest preference first):

1. TIFF files produced during the scanning and OCR of legacy literature. These files should have a resolution of at least 600 DPI to ensure that fine details in line art remain intact.

2. PDF files of (legacy) treatment. A complication may occur here when the PDF was produced using either bad-quality original scans or was highly compressed. In both cases the image quality will suffer. So high-quality PDFs are preferred.

3. Digital image files or high-quality scans of original drawings or photographs. These have the drawback that they may be missing corrections, measurement bars and/or text. Therefore the first two options are preferred.

If all of the options above are problematic or unavailable, a fourth option is to scan in the image(s) yourself using the printed publication.

What is an image?

Although images are generally thought of as being line art (including maps) or photo-graphs, some other contents of legacy Floras can also be considered

images and should be treated as such. This includes any hand drawn contents not specifically indicated as being an image, e.g. a signature. Furthermore, some elements and pages that can be considered as decorative are perhaps better treated as images than as text. In general, such contents can be considered line art.

Set up of folder structure for images

It is strongly suggested that you set up a hierarchical folder structure to organize the image files.

A sample folder structure organisation is shown in **Error! Reference source not found.**. Obviously there are more than two volumes per Flora, but these are not shown.

Although there is no technical reason to actu-ally have different folders for the three different major steps covered in this manual, doing it this way seriously reduces the risk of accidentally overwriting the original files or the files that have been processed in Photoshop. This can especially be important when running the batch process to add the metadata (in case something goes *wrong*[2]…).

The rest of the hierarchy is simply there to keep a good overview.

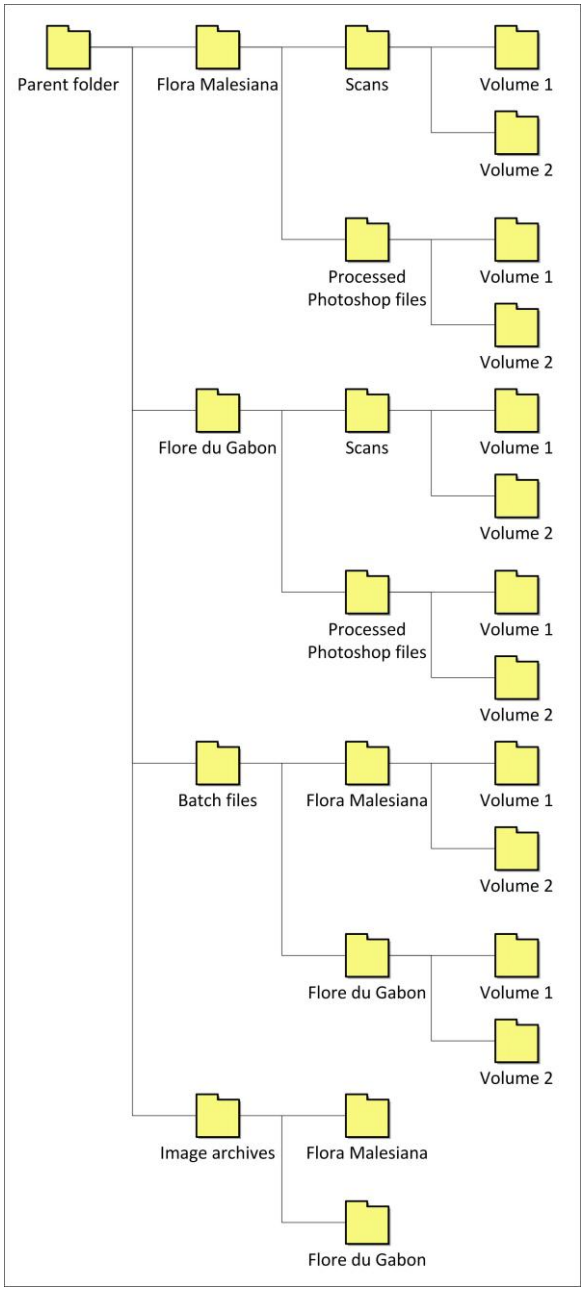


Figure 8: A suggested folder hierarchy.

[2] Things will go wrong. It's better to anticipate.

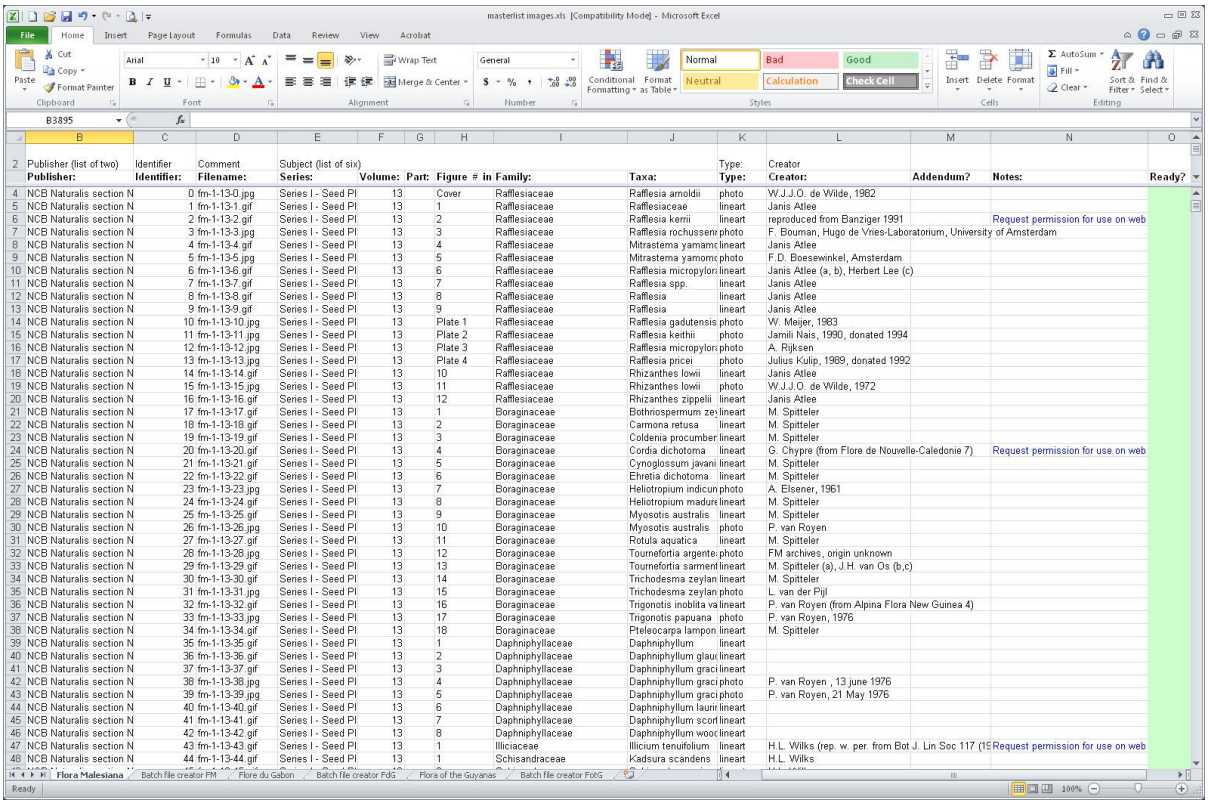
Set up of Image Administration

A local Image Administration will need to be set up to keep track of all the images and the metadata that will be added to the images later on. It is possible to use a database for this, but a large Microsoft Excel table is just as suitable.

In this table, each Flora has its own worksheet. Figure 9 shows the worksheet for Flora Malesiana, Figure 10 that of Flore du Gabon.

As can be seen in these figures, each worksheet contains fields of information on each of the images present in the taxonomic works I am working with.

Table 1 lists the common ones.



	B	C	D	E	F	G	H	I	J	K	L	M	N	O
2	Publisher (list of two)	Identifier	Comment	Subject (list of six)						Type:	Creator			
	Publisher:	**Identifier:**	**Filename:**	**Series:**	**Volume:**	**Part:**	**Figure # in**	**Family:**	**Taxa:**	**Type:**	**Creator:**	**Addendum?**	**Notes:**	**Ready?**
4	NCB Naturalis section N	0	fm-1-13-0.jpg	Series I - Seed Pl	13		Cover	Rafflesiaceae	Rafflesia arnoldii	photo	W.J.J.O. de Wilde, 1982			
5	NCB Naturalis section N	1	fm-1-13-1.gif	Series I - Seed Pl	13		1	Rafflesiaceae	Rafflesiaceae	lineart	Janis Atlee			
6	NCB Naturalis section N	2	fm-1-13-2.gif	Series I - Seed Pl	13		2	Rafflesiaceae	Rafflesia kerrii	lineart	reproduced from Banziger 1991		Request permission for use on web	
7	NCB Naturalis section N	3	fm-1-13-3.jpg	Series I - Seed Pl	13		3	Rafflesiaceae	Rafflesia rochussen	photo	F. Bouman, Hugo de Vries-Laboratorium, University		of Amsterdam	
8	NCB Naturalis section N	4	fm-1-13-4.gif	Series I - Seed Pl	13		4	Rafflesiaceae	Mitrastema yamamc	lineart	Janis Atlee			
9	NCB Naturalis section N	5	fm-1-13-5.jpg	Series I - Seed Pl	13		5	Rafflesiaceae	Mitrastema yamamc	photo	F.D. Boesewinkel, Amsterdam			
10	NCB Naturalis section N	6	fm-1-13-6.gif	Series I - Seed Pl	13		6	Rafflesiaceae	Rafflesia micropylor	lineart	Janis Atlee (a, b), Herbert Lee (c)			
11	NCB Naturalis section N	7	fm-1-13-7.gif	Series I - Seed Pl	13		7	Rafflesiaceae	Rafflesia spp.	lineart	Janis Atlee			
12	NCB Naturalis section N	8	fm-1-13-8.gif	Series I - Seed Pl	13		8	Rafflesiaceae	Rafflesia	lineart	Janis Atlee			
13	NCB Naturalis section N	9	fm-1-13-9.gif	Series I - Seed Pl	13		9	Rafflesiaceae	Rafflesia	lineart	Janis Atlee			
14	NCB Naturalis section N	10	fm-1-13-10.jpg	Series I - Seed Pl	13		Plate 1	Rafflesiaceae	Rafflesia gadutensis	photo	W. Meijer, 1963			
15	NCB Naturalis section N	11	fm-1-13-11.jpg	Series I - Seed Pl	13		Plate 2	Rafflesiaceae	Rafflesia keithii	photo	Jamili Nais, 1990, donated 1994			
16	NCB Naturalis section N	12	fm-1-13-12.jpg	Series I - Seed Pl	13		Plate 3	Rafflesiaceae	Rafflesia micropylor	photo	A. Rijksen			
17	NCB Naturalis section N	13	fm-1-13-13.jpg	Series I - Seed Pl	13		Plate 4	Rafflesiaceae	Rafflesia pricei	photo	Julius Kulip, 1989, donated 1992			
18	NCB Naturalis section N	14	fm-1-13-14.gif	Series I - Seed Pl	13		10	Rafflesiaceae	Rhizanthes lowii	lineart	Janis Atlee			
19	NCB Naturalis section N	15	fm-1-13-15.jpg	Series I - Seed Pl	13		11	Rafflesiaceae	Rhizanthes lowii	photo	W.J.J.O. de Wilde, 1972			
20	NCB Naturalis section N	16	fm-1-13-16.gif	Series I - Seed Pl	13		12	Rafflesiaceae	Rhizanthes zippelii	lineart	Janis Atlee			
21	NCB Naturalis section N	17	fm-1-13-17.gif	Series I - Seed Pl	13		1	Boraginaceae	Bothriospermum zey	lineart	M. Spitteler			
22	NCB Naturalis section N	18	fm-1-13-18.gif	Series I - Seed Pl	13		2	Boraginaceae	Carmona retusa	lineart	M. Spitteler			
23	NCB Naturalis section N	19	fm-1-13-19.gif	Series I - Seed Pl	13		3	Boraginaceae	Coldenia procumber	lineart	M. Spitteler			
24	NCB Naturalis section N	20	fm-1-13-20.gif	Series I - Seed Pl	13		4	Boraginaceae	Cordia dichotoma	lineart	G. Chypre (from Flore de Nouvelle-Caledonie 7)		Request permission for use on web	
25	NCB Naturalis section N	21	fm-1-13-21.gif	Series I - Seed Pl	13		5	Boraginaceae	Cynoglossum javani	lineart	M. Spitteler			
26	NCB Naturalis section N	22	fm-1-13-22.gif	Series I - Seed Pl	13		6	Boraginaceae	Ehretia dichotoma	lineart	M. Spitteler			
27	NCB Naturalis section N	23	fm-1-13-23.jpg	Series I - Seed Pl	13		7	Boraginaceae	Heliotropium indicun	photo	A. Elsener, 1961			
28	NCB Naturalis section N	24	fm-1-13-24.gif	Series I - Seed Pl	13		8	Boraginaceae	Heliotropium madure	lineart	M. Spitteler			
29	NCB Naturalis section N	25	fm-1-13-25.gif	Series I - Seed Pl	13		9	Boraginaceae	Myosotis australis	lineart	M. Spitteler			
30	NCB Naturalis section N	26	fm-1-13-26.jpg	Series I - Seed Pl	13		10	Boraginaceae	Myosotis australis	photo	P. van Royen			
31	NCB Naturalis section N	27	fm-1-13-27.gif	Series I - Seed Pl	13		11	Boraginaceae	Rotula aquatica	lineart	M. Spitteler			
32	NCB Naturalis section N	28	fm-1-13-28.jpg	Series I - Seed Pl	13		12	Boraginaceae	Tournefortia argente	photo	FM archives, origin unknown			
33	NCB Naturalis section N	29	fm-1-13-29.gif	Series I - Seed Pl	13		13	Boraginaceae	Tournefortia sarment	lineart	M. Spitteler (a), J.H. van Os (b,c)			
34	NCB Naturalis section N	30	fm-1-13-30.gif	Series I - Seed Pl	13		14	Boraginaceae	Trichodesma zeylan	lineart	M. Spitteler			
35	NCB Naturalis section N	31	fm-1-13-31.jpg	Series I - Seed Pl	13		15	Boraginaceae	Trichodesma zeylan	photo	L. van der Pijl			
36	NCB Naturalis section N	32	fm-1-13-32.gif	Series I - Seed Pl	13		16	Boraginaceae	Trigonotis inoblita va	lineart	P. van Royen (from Alpina Flora New Guinea 4)			
37	NCB Naturalis section N	33	fm-1-13-33.jpg	Series I - Seed Pl	13		17	Boraginaceae	Trigonotis papuana	photo	P. van Royen, 1976			
38	NCB Naturalis section N	34	fm-1-13-34.gif	Series I - Seed Pl	13		18	Boraginaceae	Pteleocarpa lampon	lineart	M. Spitteler			
39	NCB Naturalis section N	35	fm-1-13-35.gif	Series I - Seed Pl	13		1	Daphniphyllaceae	Daphniphyllum	lineart				
40	NCB Naturalis section N	36	fm-1-13-36.gif	Series I - Seed Pl	13		2	Daphniphyllaceae	Daphniphyllum glau	lineart				
41	NCB Naturalis section N	37	fm-1-13-37.gif	Series I - Seed Pl	13		3	Daphniphyllaceae	Daphniphyllum graci	lineart				
42	NCB Naturalis section N	38	fm-1-13-38.jpg	Series I - Seed Pl	13		4	Daphniphyllaceae	Daphniphyllum graci	photo	P. van Royen , 13 june 1976			
43	NCB Naturalis section N	39	fm-1-13-39.jpg	Series I - Seed Pl	13		5	Daphniphyllaceae	Daphniphyllum graci	photo	P. van Royen, 21 May 1976			
44	NCB Naturalis section N	40	fm-1-13-40.gif	Series I - Seed Pl	13		6	Daphniphyllaceae	Daphniphyllum laurir	lineart				
45	NCB Naturalis section N	41	fm-1-13-41.gif	Series I - Seed Pl	13		7	Daphniphyllaceae	Daphniphyllum scort	lineart				
46	NCB Naturalis section N	42	fm-1-13-42.gif	Series I - Seed Pl	13		8	Daphniphyllaceae	Daphniphyllum wooc	lineart				
47	NCB Naturalis section N	43	fm-1-13-43.gif	Series I - Seed Pl	13		1	Illiciaceae	Illicium tenuifolium	lineart	H.L. Wilks (rep. w. per. from Bot J. Lin Soc 117 (1		Request permission for use on web	
48	NCB Naturalis section N	44	fm-1-13-44.gif	Series I - Seed Pl	13		1	Schisandraceae	Kadsura scandens	lineart	H.L. Wilks			

Figure 9: Image Administration worksheet for Flora Malesiana.

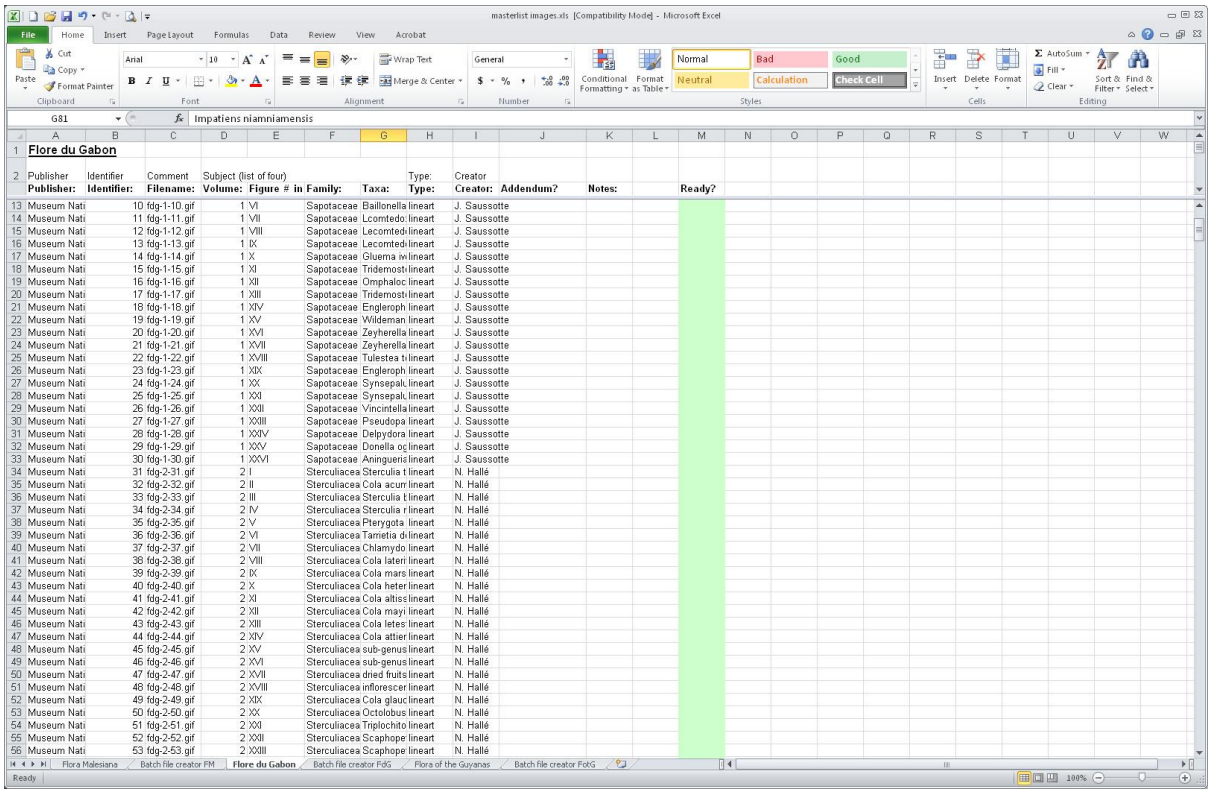


Figure 10: Image Administration worksheet for Flore du Gabon.

Table 1: Information fields in Image Administration.

Field	Explanation
Publisher	Who published the taxonomic work?
Identifier	Unique ID identifying each image in a flora.
File name	The file name the image will be getting. See “Naming image files” below.
Volume number	The volume number of the printed work.
Figure # in treatment	The figure number printed in the legacy treatment, including prefix if the prefix is not “Figure” or “Fig.”
Family	The family the taxa depicted belong to.
Taxa	The taxa depicted.
Type	The image type; with choice of line art, photo, other.
Creator	The creator of the figure, where it comes from.
Addendum?	Used to indicate whether an image belongs to an addendum, erratum or corrigendum (addition, error, or correction), which are often present at the back of legacy Flora volumes to be taken into account in later revisions. In some cases these include replacement figures or photographs.
Notes	Used for notes from the person who processes the images, e.g. whether additional permission should be asked to use images.

All of this information will later be added to each of the image files as metadata. The infor-mation recorded may include additional fields if required, as shown in Figure 9 for Flora Malesiana. This Flora consists of two series, one treating angiosperms, the other ferns, so there is a column to indicate the series. Furthermore, some volumes are split into parts, meaning another additional column.

The “Ready?” column is not actually used to record information, but is used to indicate whether or not an image has been fully prepared for the web (by changing the color to green).

Each Flora has an additional worksheet that is used to facilitate creating Batch files (called "Batch file creator <insert Flora abbreviation here>"), that is listed at the bottom of the screen (Figure 11). How this functions is described elsewhere.

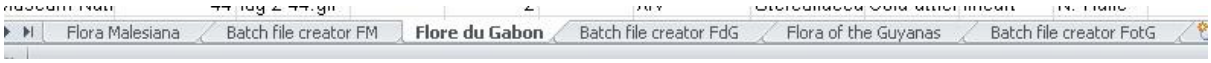


Figure 11: Worksheets in Image Administration

Tip: Figure 9 and Figure 10 also show that Microsoft Excel’s “Split” view is in use to keep showing the table headings at all times. This view can be switched on by choosing “Split” or “Split Windows” (depending on the version of Excel) in Excel’s “View” menu. Two separators will show up in the middle of the win-dow (Figure 12). The vertical separator can be removed by dragging it entirely to the left. Now scroll entirely up in the upper half of the window, and then move the horizontal separator to where you want it. When the Excel file is saved, the separators are saved too.

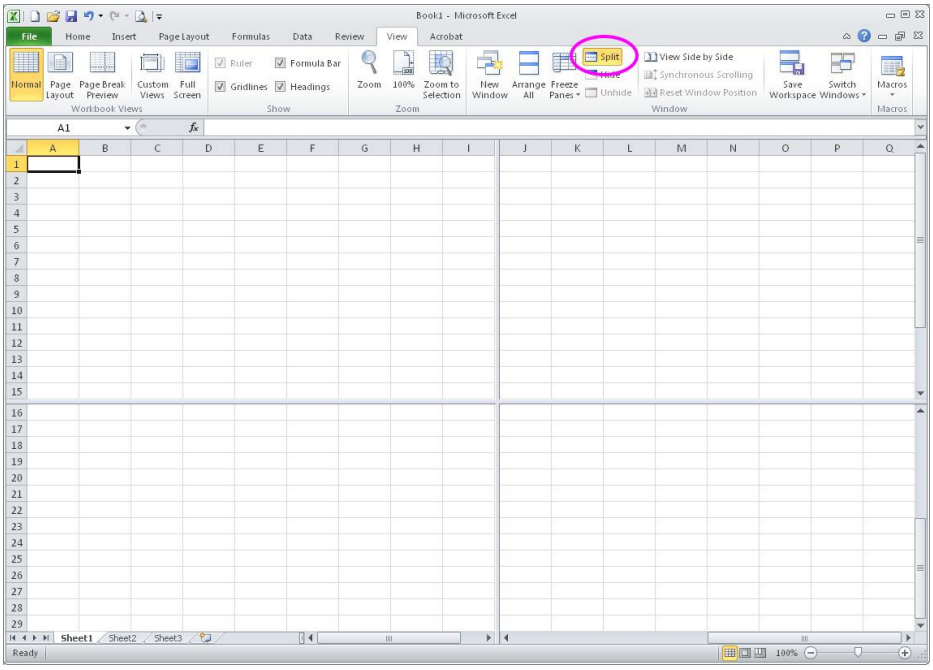


Figure 12: How to activate the split view in Microsoft Excel.

Naming image files

The naming convention used for naming image files does not depend on the taxon shown in the image. This is a deliberate choice. Naming each image after its contents would require renaming the files any time a taxonomic name is changed. Instead, each image receives a name indicating from what taxonomic work it comes, using the Flora title's initials, the Flora subseries (if required), the Flora volume and the unique identifier given to that image within the Image Administration. For example, an image may receive the name "fm-2-3-3803.jpg", which means it is from **F**lora **M**alesiana Series **II** Volume **3**, and has ID **3803**. Being a JPEG image it also likely is a photograph. Likewise, "fdg-2-46.gif" is from **F**lore **d**u **G**abon Volume **2**, with ID **46**, and is likely line art since it is a GIF.

The filename is also part of the metadata that gets added to the image, because that means that if the file is accidentally renamed the filename can be retrieved from within the image file itself with only a little trouble.

Image file formats

In the previous section two image file formats have been mentioned: GIF and JPEG. These formats were chosen for three reasons:

- They are well-supported in almost all web browsers since about two decades.
- When used properly, they allow for very small file sizes without excess loss of quality.
- They are sufficient for the types of images we have to deal with.

The second point may require some additional explanation. You might think "But we have great internet speeds and technology that allows for fast transmission of large quantities of data!", but this is partially wrong. Such facilities are present *wherever the infrastructure can support it*, in developed countries and urbanized areas in general. However, even within Europe and North America there remain rural areas that are stuck on relatively slow dial-up internet connections, instead of large-capacity cable or DSL connections. Institutional con-nections will generally be fast, even in developing countries, but the same cannot be said of private connections or connections in the field. Furthermore, people might not be running the latest software, which can also lead to problems. Why not help these people out by not serv-ing them huge images?

The images used in legacy Floras come in three flavours: 1) line art, 2) black and white photographs with a varying degree of quality (usually quite bad), and 3) colour photographs. The GIF (Graphics Interchange Format) is excellent for line art, and allows image size to be reduced by 50-75% without any obvious loss of quality. Scanned in photographs from printed legacy publications generally suffer badly from Moiré patterns (http://en.wikipedia.org/wiki/Moiré_pattern) caused by the rasterisation during the original printing process. Furthermore, black and white photographs taken in the field especially suffer from issues related to exposure or

unsuitable film speeds that are exacerbated by print quality issues. These artefacts are much worse than any artefacts caused by JPEG lossy compression.

Image Administration tasks prior to image processing

Inventarisation of images to be processed

Before even starting to process images, you should have a look through the Flora volume(s) you will be processing, and count the amount of figures. You then should copy the figure numbers into the appropriate column in the Image Administration. Once you know how many figures you will be processing, you can fill out as much of the information as possible for each image. You can copy the taxa shown out of the figure legends into the taxa column. If the creator of the image is indicated, you should write it down too – sometimes the people who drew the images are only mentioned at the start or end of a volume, separately from the actual images. You should also indicate whether permission should be asked to use an image. This is usually the case when the image was not produced in-house; the permission granted often only applies to one medium (print).

It is possible to do this work fully digitally, but it can be more practical to keep a hardcopy of the volume you are working on nearby, and simply work your way through the paper volume while filling out the Excel Image Administration worksheet. This is somewhat faster than a fully digital workflow.

In general, it is recommended to not fill out the file name until the moment you are actually going to process the image in Photoshop. This helps to keep track of which images have already been processed in Photoshop and which ones have not. It also reduces the chance of entering the wrong file extension into the Excel sheet, which, if not caught, will cause the metadata addition process to fail for that file. There is a column called "Ready?" that can be filled out once the entire conversion process, including metadata addition, has been completed.

Image processing in Adobe Photoshop

In this part it is explained how to process the images in Photoshop, including eventual improvements to the quality, to prepare them for the next step: adding the metadata.

Opening files

Files can be opened the usual way, via the “Open…” option in the “File”-menu. For PDFs, see below.

Warning!!! If you open the files directly from the folders with legacy taxonomic work scans, it is essential not to accidentally save the file under the same name while processing it in Photoshop, because then the original file will be re-placed (and you *might* not want *that*).

Opening a PDF

To open a PDF, use the “Open…” option under the “File”-menu as usual, select the PDF you want to open, and click Open. You then get the “Import PDF” window (Figure 13).

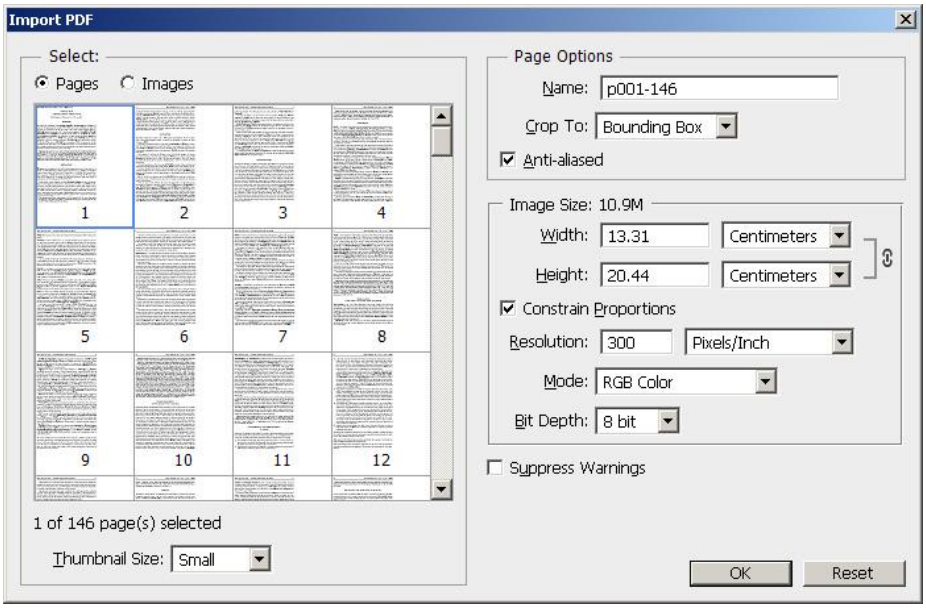


Figure 13: Import PDF window in Photoshop.

Select all of the pages that feature images and click OK.

Tip: To select multiple pages, click the first page, then press and hold the CTRL key on the keyboard while clicking the remaining pages. You can release the CTRL key to scroll down, but press it down again before you start selecting more pages, otherwise the pages you already selected will be deselected. An example is shown in Figure 14.

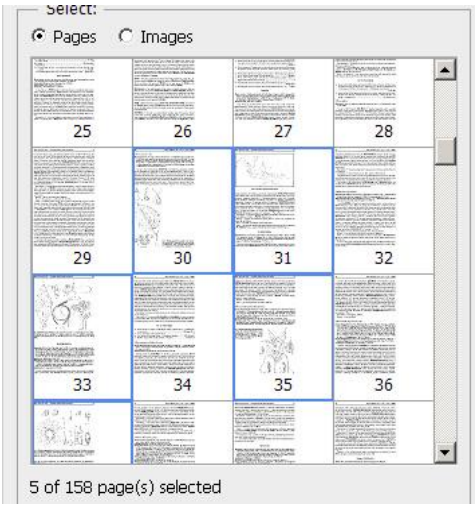


Figure 14: Multiple pages have been selected.

Figure 15 shows a page from a PDF after it has been opened. The back-ground shows a checkerboard pattern, which means it is transparent. You do not need to fix this at this time; it will be fixed when saving the image file later on. You can now carry on with processing the images from the PDF like any other image. If the procedure differs slightly for an image coming from a PDF, this will be indicated.

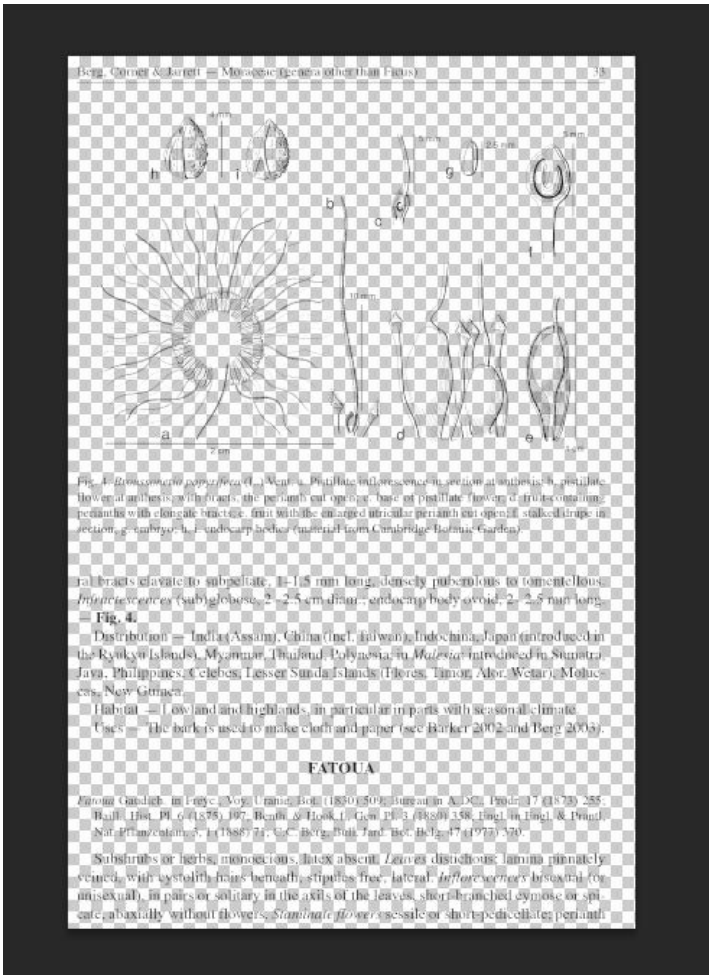


Figure 15: A PDF page loaded in Photoshop.

How to process the image files

The Undo and Step Backwards functions

You will likely make errors now and then while processing image files. Photoshop has two functions that allow you to undo actions: 1) the "Undo" option, and 2) the "Step Backwards" option, both under the "Edit"-menu. "Undo" undoes your last action(s), while "Step Backwards" allows you to revert more actions.

Image mode

The first step is changing the Image Mode, using the options under “Mode” in the “Image”-menu. This should be either Grayscale (for line art or black and white photographs) or RGB Color (for colour photographs). A scan of a legacy publication will likely be a Bitmap (Figure 16) and should therefore be changed. Recording this as a Photoshop Action is useful.

Warning!!! Not changing the Image Mode from Bitmap to Grayscale or RGB Color will lead to considerable degradation of the image quality when resizing later on. Furthermore, certain image modifications are not possible when the image mode is still set to Bitmap.

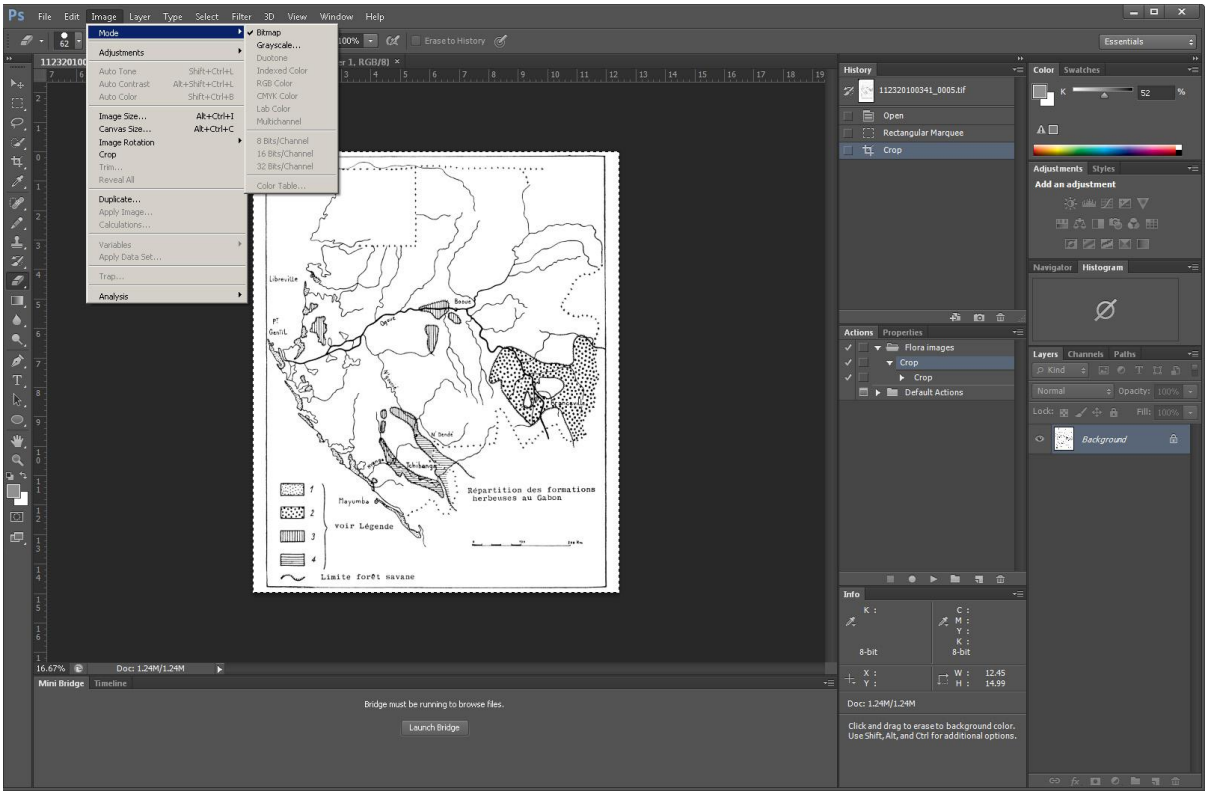


Figure 16: Image Mode. Here it is set to Bitmap, and should be changed to Grayscale.

PDF: PDF pages likely are already in RGB Color. You can change this to Grayscale, but it is not required.

Image orientation

If the image is not oriented correctly, this should be corrected. Use the options under “Rotate” in the “Image”-menu for this (Figure 17). It is suggested that you recording three Photoshop Actions here:

- Rotate 180 degrees.
- Rotate 90 degrees left (CCW).

- Rotate 90 degrees right (CW).

  Note: Unfortunately, it is not possible to record a Photoshop Action to rotate objects arbitrarily, and still be able to specify how many degrees the object should be rotated (the number of degrees is recorded as part of the Action), so to rotate arbitrarily you will still have to use the menu option manually.

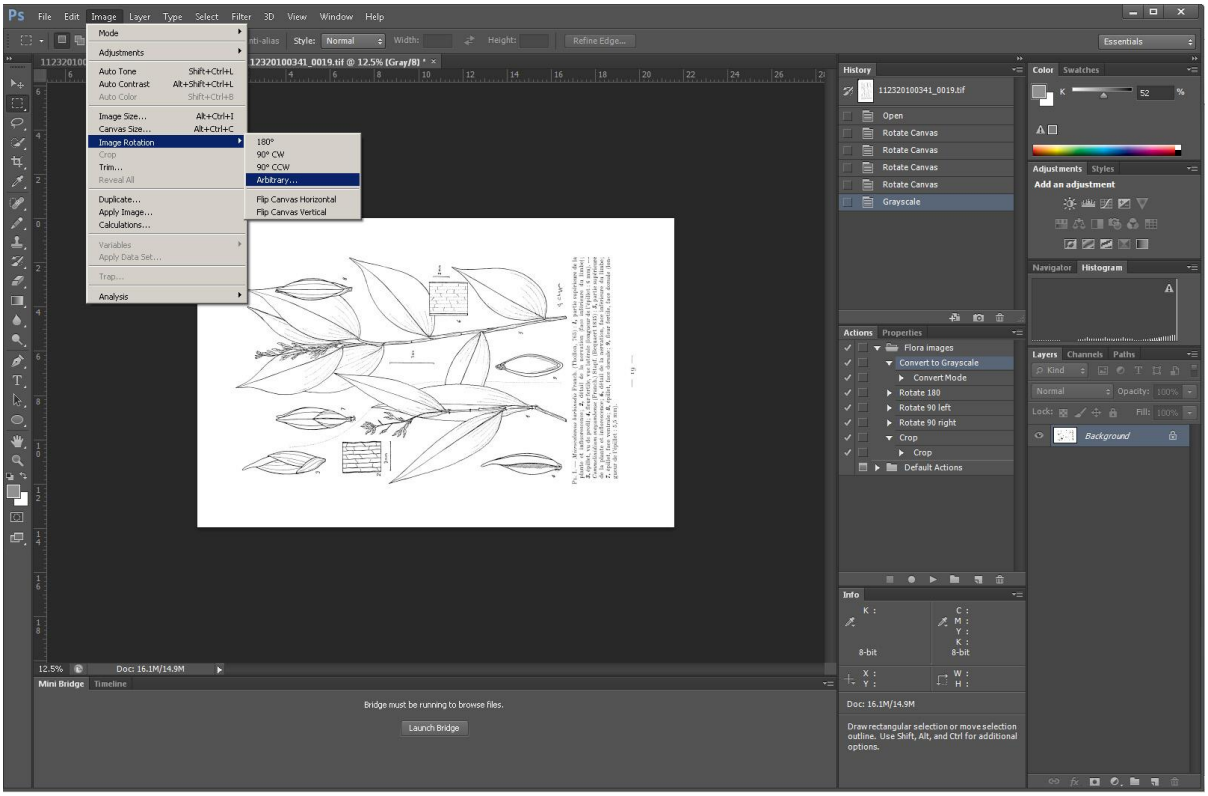


Figure 17: Rotation options, and an image that will need to be rotated 90 degrees clockwise.

Cropping

The image has to be cropped to remove anything unneeded. To do this, in the toolbar at the left select the Rectangular Marquee Tool, and drag over the area to crop (Figure 18). As shown, you only select the actual figure – figure legends and anything else on the page that isn't required can be cropped off.

Then you go to the "Image"-menu and select "Crop". The result is shown in Figure 19. It is useful to record this as a Photoshop Action for future use.

In some cases there will be multiple figures on a single page, each with their own figure legend. These will all need to be cropped (and saved) separately. After saving the first image, you can go back to the original image using the "Step Backward" menu option under the "Edit"-menu. An alternative is to open the original image again.

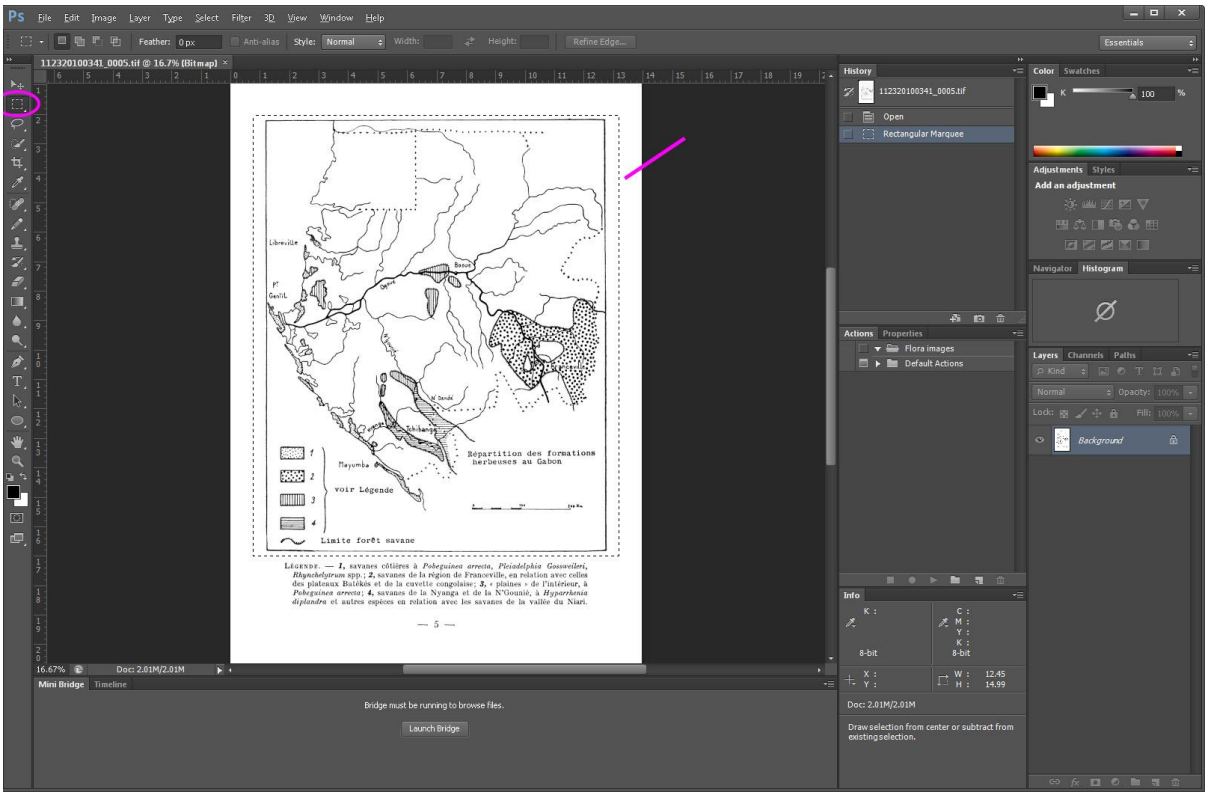


Figure 18: Cropping an image. The Rectangular Marquee Tool is circled, and the selection area is indicated.

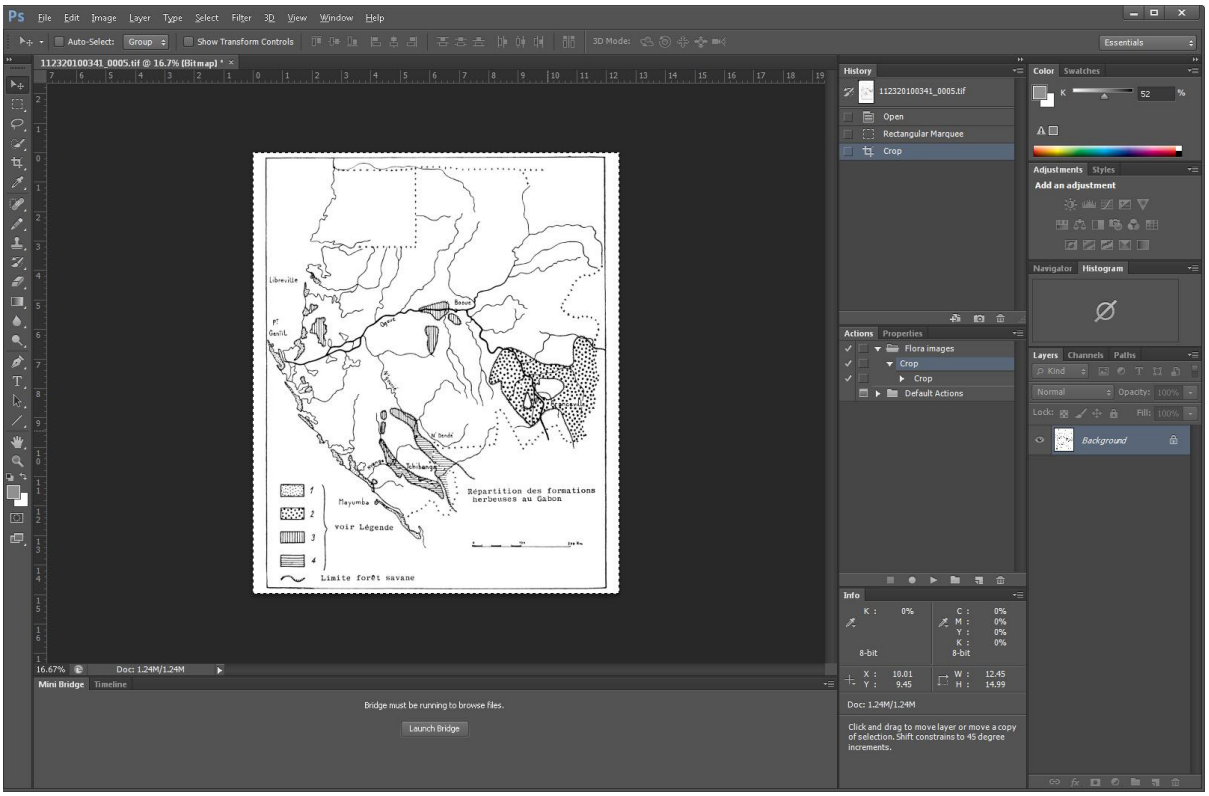


Figure 19: Cropped image.

Removing unwanted text or other problems

Sometimes the cropped image will include excess text or other problems, such as stains or dirt. You can use the Eraser Tool to remove such issues (Figure 20).

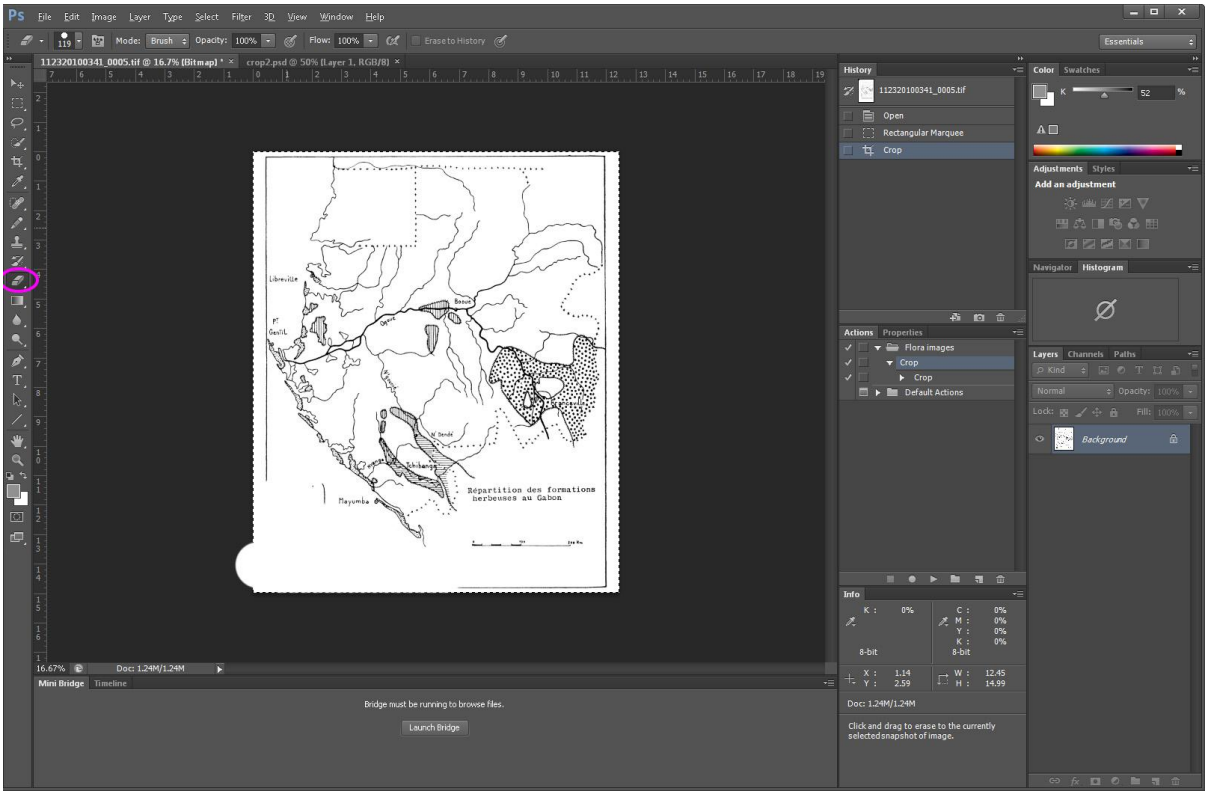


Figure 20: Eraser Tool.

Image quality improvements

Line art generally does not need to be improved, but most (black and white) photographs will be either too light, too dark, or have a lack of contrast. This can somewhat be corrected by going to the “Image”-menu, choosing “Adjustments” and then “Brightness/Contrast” (Figure 21). The contrast can usually be increased by 10-40%, while the brightness should be adjusted depending on the brightness (or lack thereof) of the actual image. You can best use brightness adjustment steps of 10-20%. Again, recording this as one or more Photoshop Actions is useful.

(continued next page)

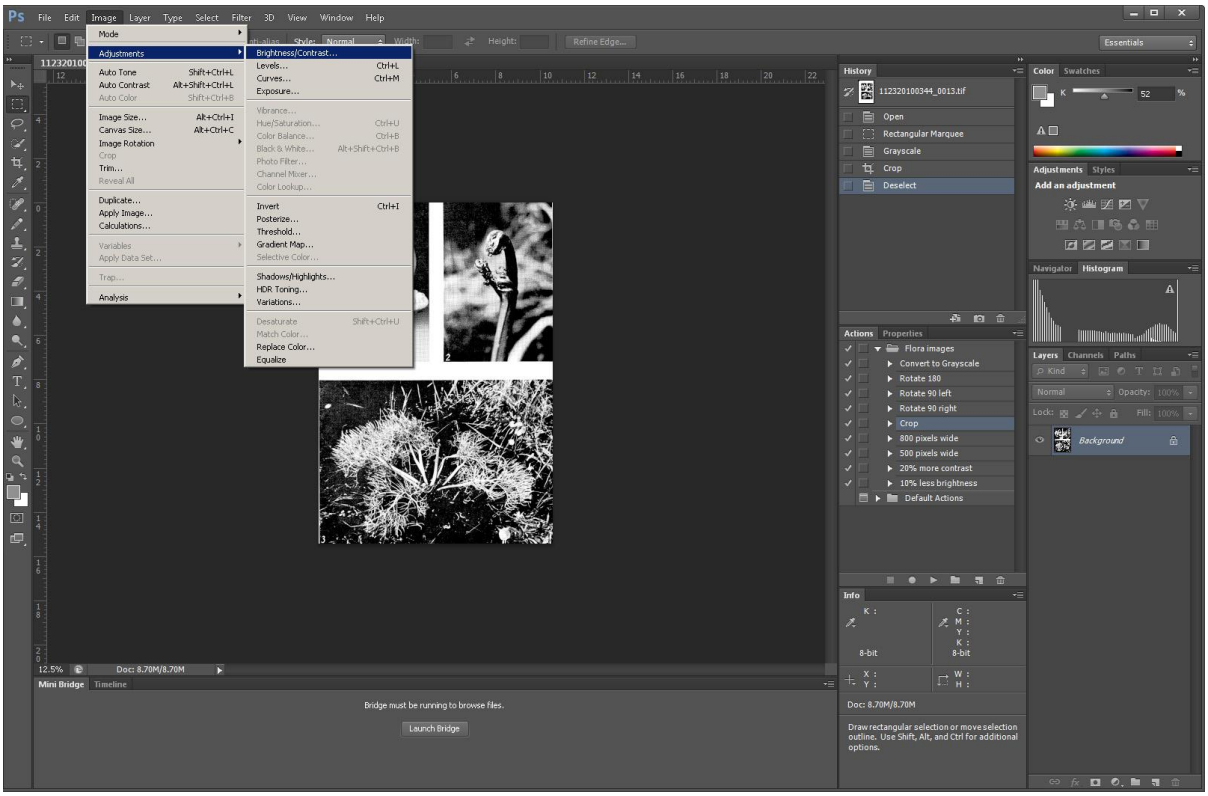


Figure 21: Adjusting the brightness and/or the contrast.

Resizing

Images do not need to be 3,000-4,000 pixels wide for use in an e-Flora, nor do they need to be that width for use in a printed version of that e-Flora. Therefore you can resize them to more acceptable sizes, which also results in much smaller file sizes. A width of 800 pixels for images as wide as the Flora page they are on, and 500 pixels for images half the width of a page usually results in images that suffer no visible degradation when printed larger than life. To resize, you go to the "Image"-menu, and choose "Image Size…". This results in the "Image Size" window (Figure 22), where the new image width can be specified. It is very important that "Constraint Proportions" has a checkmark; otherwise the image may be distorted when it is resized. "Scale Styles" and "Resample Images" should also have a checkmark. Resizing should also be recorded as Photoshop Actions.

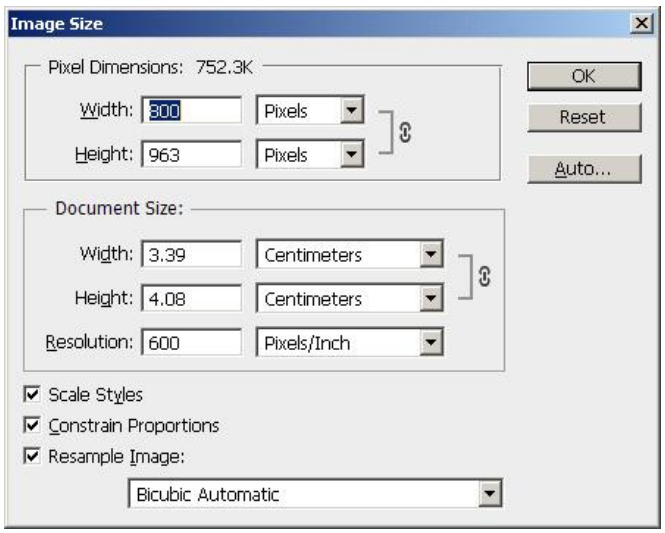


Figure 22: Image Size window.

Saving

To save the file, you can use the regular “Save as…” option in the “File”-menu, but you can better use the “Save for Web…” option in the same menu (Figure 23). This option has the advantage that images are optimized for use on the internet. You can fine-tune several options to have a good image quality and small file size. Another advantage of the “Save for Web…” option is that the original file is not modified.

Figure 24 shows the actual “Save for Web” window with all the options. It is important to select the correct file format in the top right corner (circled): “GIF” for line art, “JPEG” for photos. You can play a bit with the various options and observe any degraded image quality in the preview window (the preview window tab should be set to "Optimized").

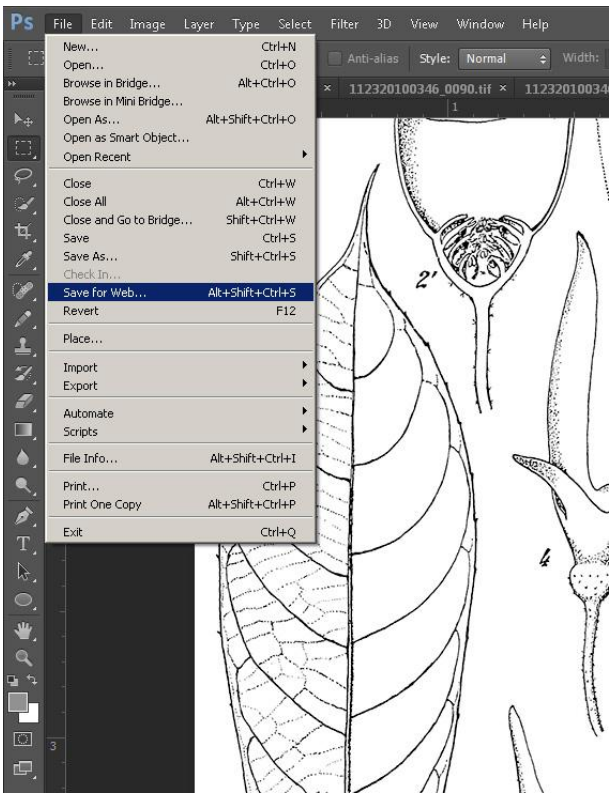


Figure 23: “Save for Web…” option in file menu.

PDF: When working with images extracted from a PDF, you should remove the checkmark in front of “Transparency” to get a white background.

(continued next page)

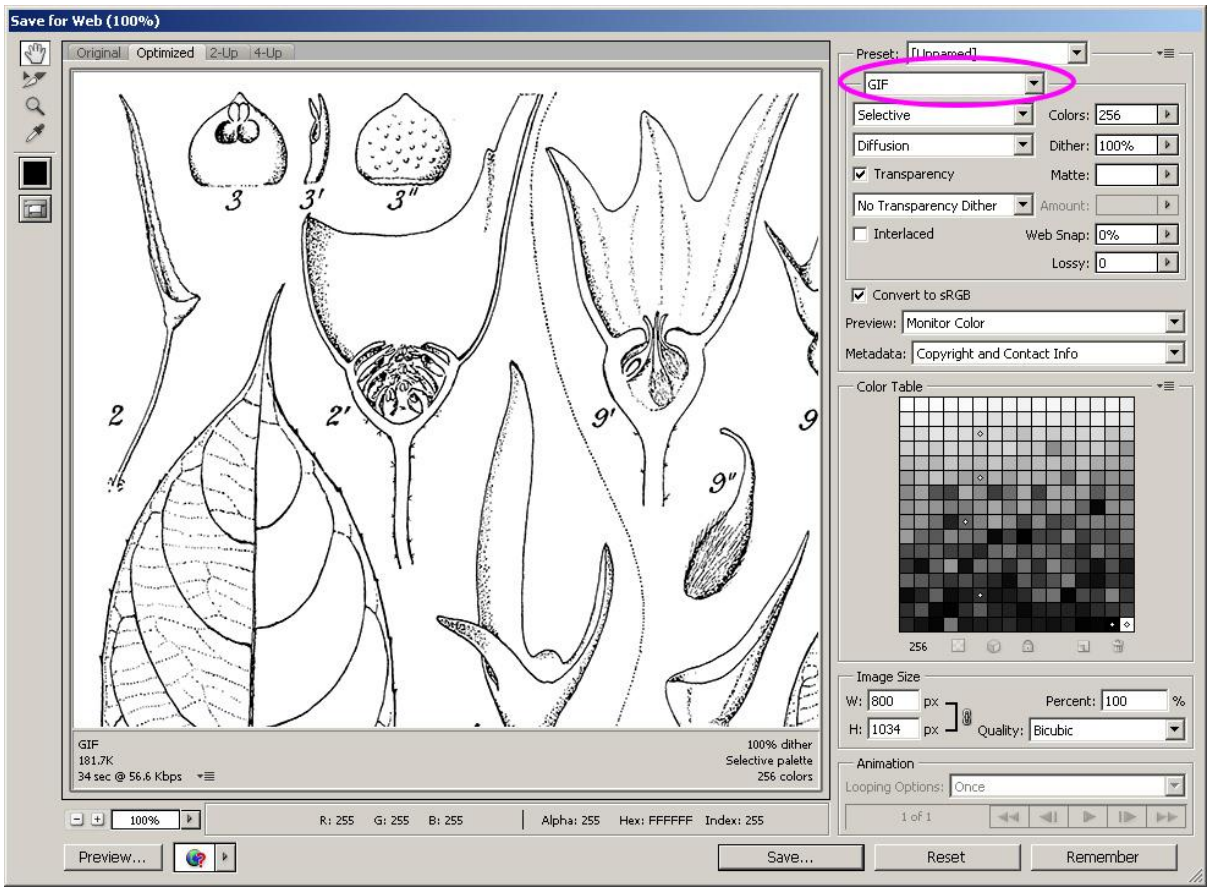


Figure 24: "Save for web…" window showing various options.

After clicking on "Save…", you can choose where to save the image file.

> **Note:** It is important that the image files' file names are exactly the same in the Image Administration and in reality. Otherwise adding the metadata will fail.

You should save the image files to their volume's folder for images that have been processed in Photoshop.

Metadata

Warning!!! DO NOT try to add the metadata to the images prior to actually processing them in Photoshop! Not only will this put a very heavy load on your computer while running the batch process due to much larger image file sizes, there also is a chance that the metadata will get lost during Photoshop work.

Introduction

EXIFTool and batch files

EXIFTool is a free command line tool (application; "App") that can be used to read, write and modify the meta-data that are embedded into image files of various formats. It can be downloaded from http://www.sno.phy.queensu.ca/~phil/exiftool/ (you need the Windows executable, assuming you are on Windows) and is extensively documented on that website. Unfortunately, the documentation requires a fair amount of technical knowledge to read, and the syntax for commands is not exactly simple. Furthermore, used from the Windows Command Prompt only one file can be processed at once.

This is where batch files come in. A batch file is a text file containing commands that are executed by the command interpreter (in this case Windows Command Prompt) when the batch file is run, either by double-clicking it from within Windows or by running it directly from the Command Prompt.

You will use batch files to add metadata to multiple files in one go, by creating batch files that contain all of the required EXIFTool commands for each of the image files to which metadata should be added.

Tip: Although EXIFTool's syntax and use is somewhat complicated, you should still read the documentation anyway, otherwise you will not understand how EXIFTool's syntax works and how to add additional metadata if required.

What is the command line?

The Command Prompt, or MS-DOS Prompt (in older versions of Windows), or Shell (in UNIX-likes) is a text-only user interface from which you can run programs, do system tasks, and much more. In Windows it generally is only encountered when you have to perform more advanced tasks.

Using Microsoft Excel for easy production of batch files

Because writing batch files by hand remains fairly complicated and error-prone, you will use Microsoft Excel to create such files. As mentioned before, the Image Administration should feature an additional worksheet for this purpose for each of the Floras you work on (Figure 11). Figure 25 shows how the “Batch file creator”-worksheet (selected by clicking on the appropriate worksheet tab in Microsoft Excel; see Figure 11) looks for Flora Malesiana.

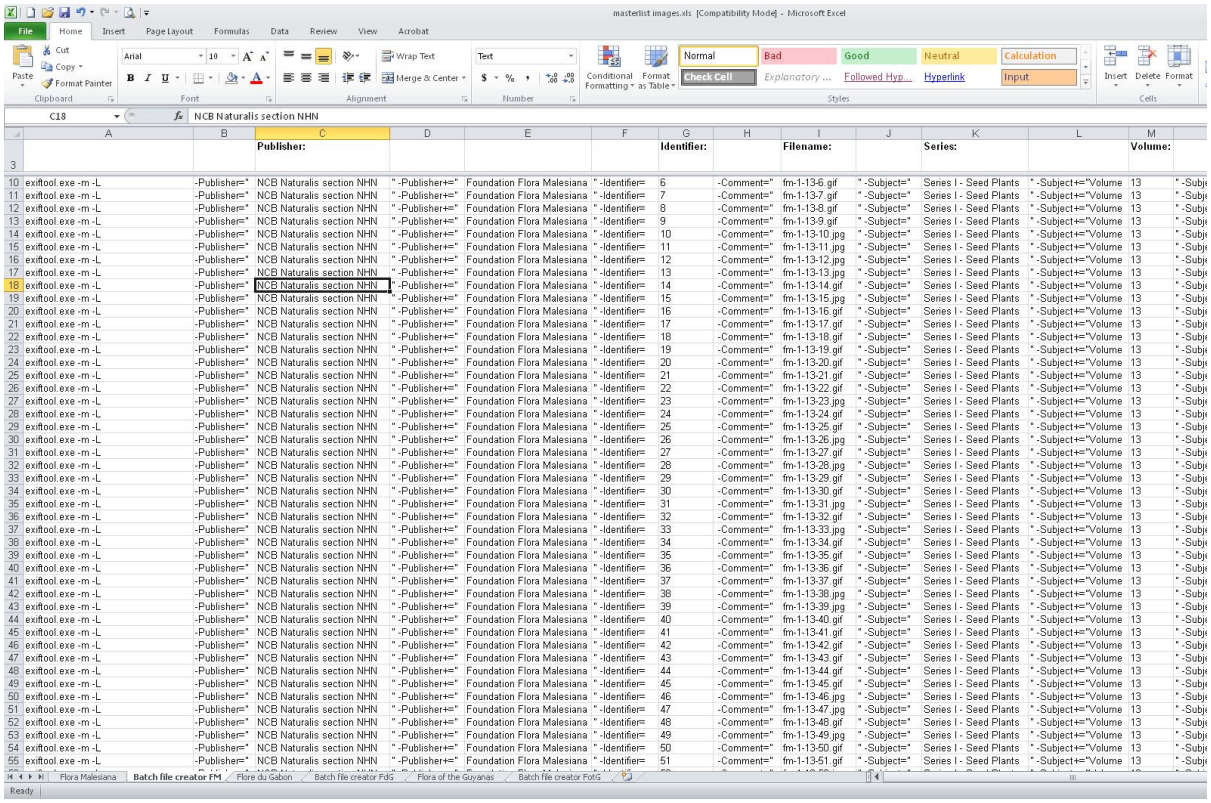


Figure 25: The “Batch file creator” for Flora Malesiana.

Tip: Some handy shortcuts to use in Excel:

- Copy: CTRL-c
- Paste: CTRL-v
- Selecting multiple cells: either click and drag, or click the first cell, hold down the Shift key, and click the last cell.

As can be seen in the figure, certain columns contain part of the syntax used in EXIFTool commands, while other columns contain the metadata that will be added to the image files. The columns that contain metadata have headings that correspond to the headings used in the actual Image Administration worksheet for that Flora.

The “Batch file creator” worksheet is subsequently filled out and converted to a text file. Then some very simple modifications are made to that text file to change it into a batch file. This procedure is described in detail in the next section.

Creating a batch file

Figure 26 shows the “Batch file creator” for Flore du Gabon. This may be similar to what you might start with. Grey cells are cells that have already been processed, and the line you are starting at is indicated with pink arrows.

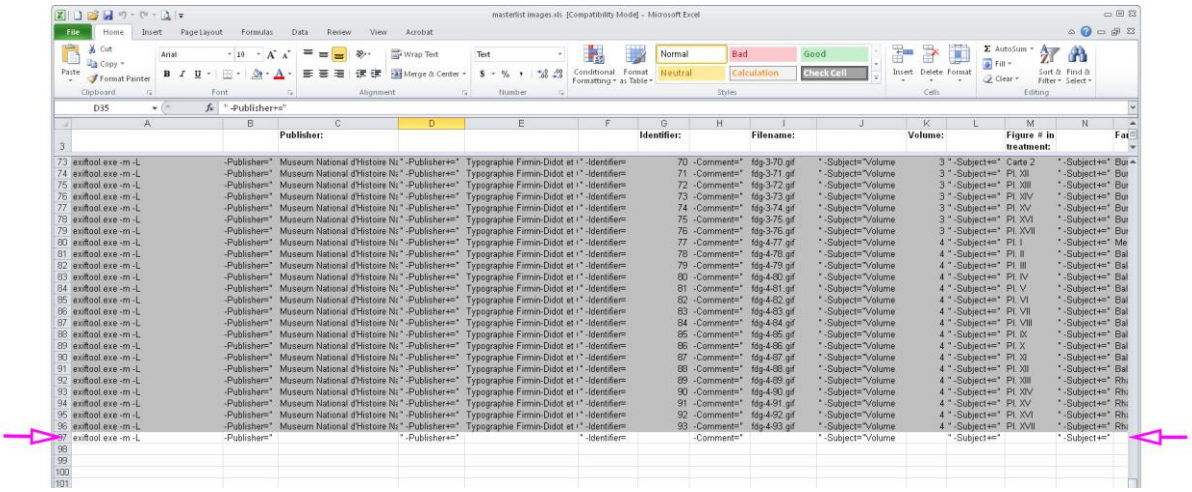


Figure 26: The “Batch file creator” for Flore du Gabon, prior to filling out. Contents with a gray background have been processed earlier. The arrows indicate the line that will be started at.

Count the number of images you will be adding in the Image Administration, then select the cells containing only EXIFTool syntax in the “Batch file creator” worksheet for Flore du Gabon. Figure 27 shows how to do this. Go to the bottom right corner of the selected cells (arrow 1 in Figure 27; the mouse pointer should change into a thin cross), press the left mouse button, and drag down until you have covered enough cells for the number of images (arrow 2 in Figure 27). The selected cells will be duplicated in the cells below the first row.

Then the contents of each column can be copied from the actual Image Administration work-sheet to the “Batch file creator” worksheet. Note that the file name goes into two columns (Figure 28). Be sure to save the file once you are done.

(continued next page)

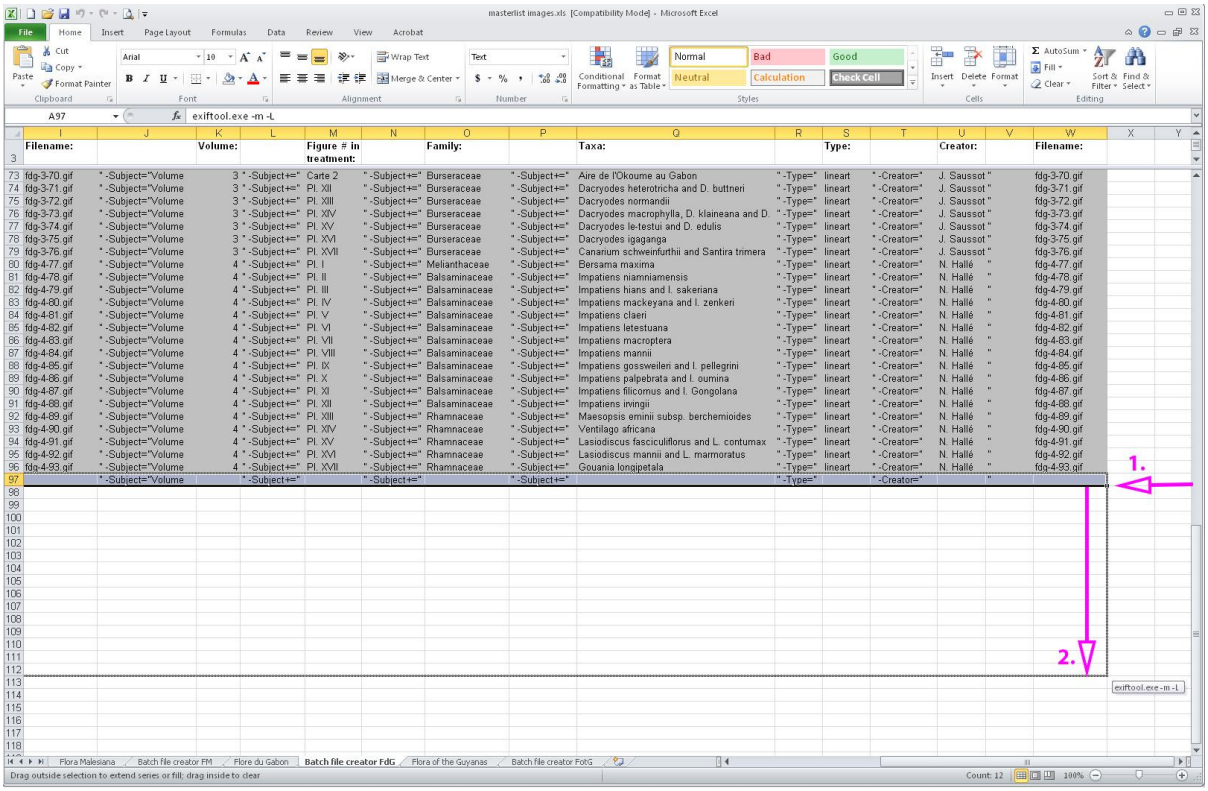


Figure 27: Replicating text into multiple lines in Microsoft Excel. See text for details.

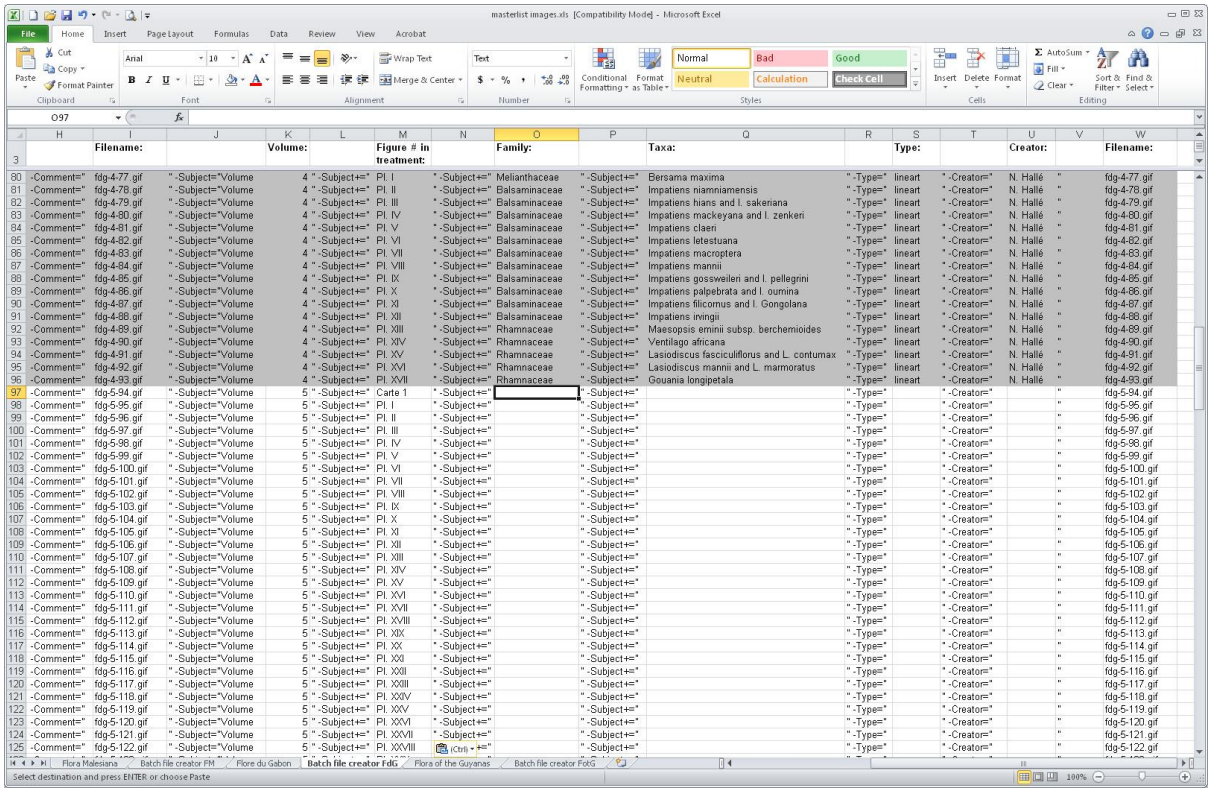


Figure 28: Copying contents from the actual image administration to the "Batch file creator".

Now select all of the cells containing EXIFTool syntax and metadata, and copy them to a blank new worksheet. Go to this worksheet.

Check whether there are empty cells in the new worksheet. These need to be deleted together with the cell that directly precedes them, to avoid complications later on. Figure 29 shows a few examples. Select the cells to delete, right-click them, and select “Delete” (just hitting the delete key on the keyboard will only clear the contents of the cells, not delete them). In the window that then shows up, select “Shift cells left” and click "OK".

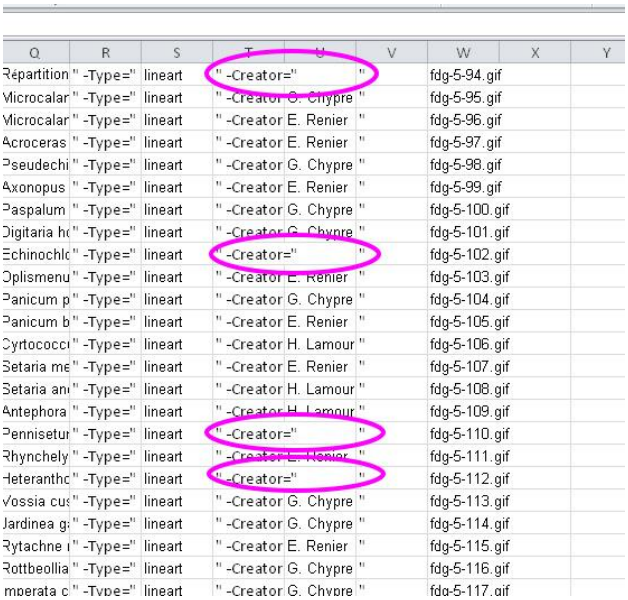


Figure 29: Empty cells that need to be removed.

Now go to “File”, “Save as”, and select the folder you use for your Batch files. Give the file a name, for example a name reminding you of the flora series and volumes the image files belong to, and **be sure** to select “Text (Tab delimited) (*.txt)” as the file type (Figure 30).

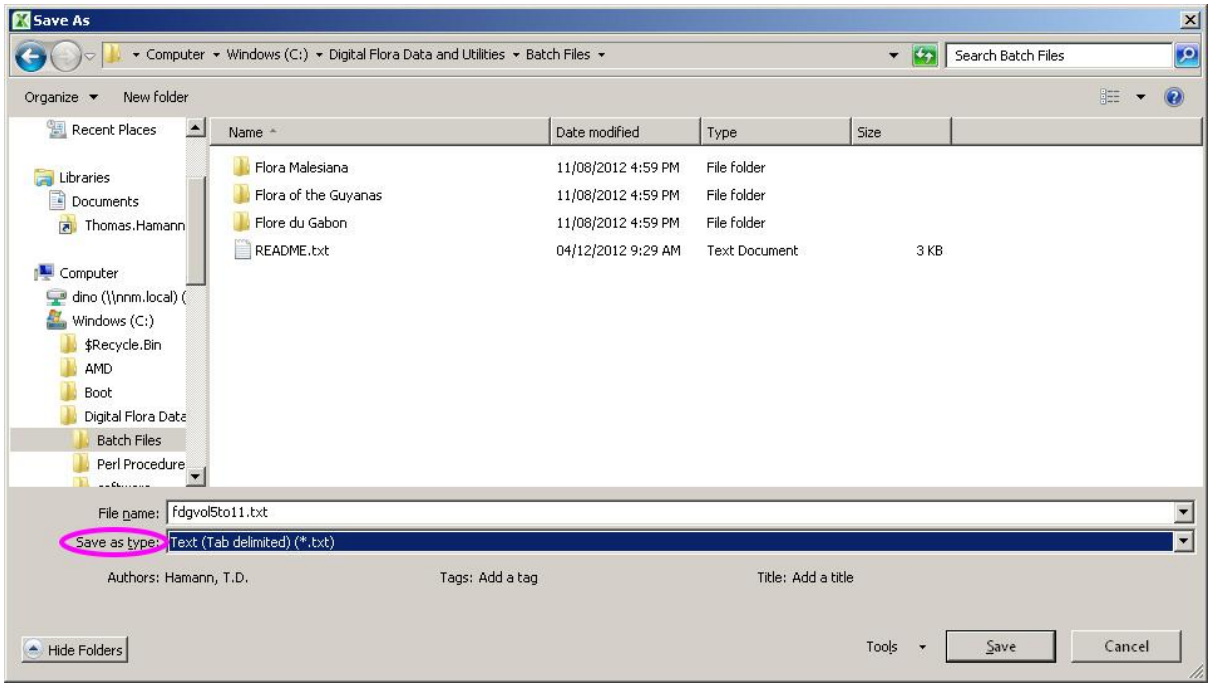


Figure 30: Saving the Excel worksheet as a tab-delimited text file.

Click on “Save”. Excel may give a warning about the number of worksheets; click “OK”, and another warning about unsupported features; click “Yes”. Close the worksheet you just saved.

Start Notepad++, and open the text file you just created. You should see something similar to Figure 31.

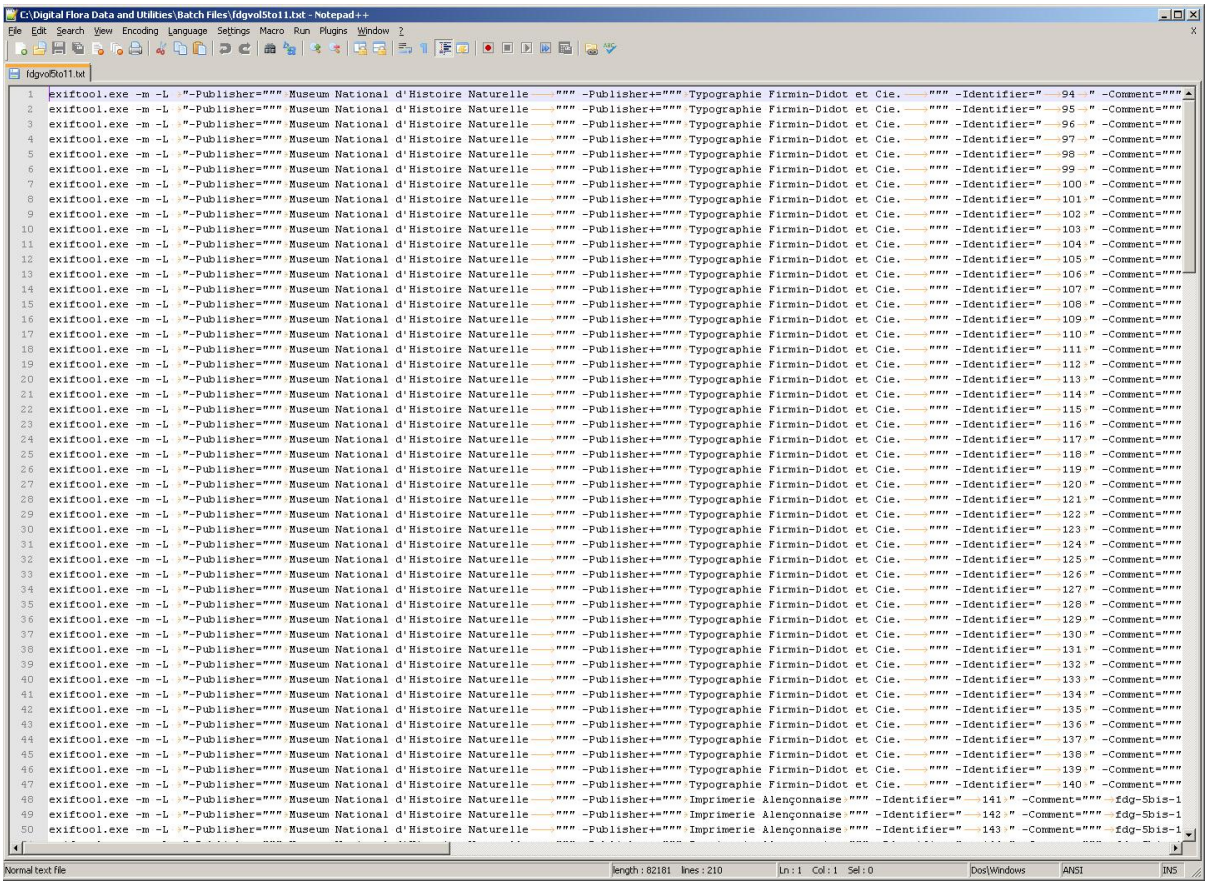


Figure 31: The future batch file in Notepad++, with Excel-inserted additional quotation marks.

As you can see, Microsoft Excel unfortunately inserts additional quotation marks in the file when saving as tab-delimited text - this appears to be a feature/bug in Excel. So these need to be removed by using Notepad++’s “Replace” function. This can be accessed via the “Search”-menu (“Replace…”) or more simply by pressing the "CTRL" and "H" keys on your keyboard at once. Figure 32 shows the Replace window, and the options suggested. “Search Mode” should be set to “Extended”.

(continued next page)

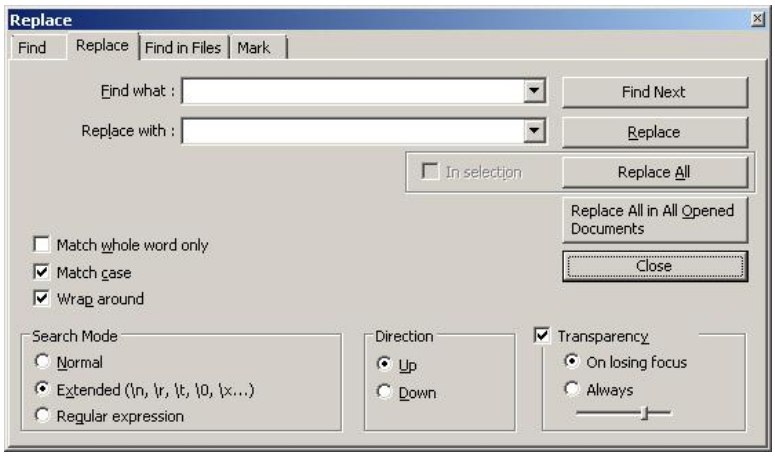


Figure 32: Notepad++'s Replace window.

The order of the following steps is important. The text to be inserted has been highlighted, so you can see when spaces are parts of the pattern. You can always use "Replace All".

1) Replace """ (three quotation marks) by " (one quotation mark).
2) Replace "-Publisher by -Publisher (note lack of space in former).
3) Replace \t by nothing at all.
4) Replace "" (two quotation marks) by " (one quotation mark).
5) Replace "fdg by fdg (replace fdg by the initials of your flora that you used for your filenames).
6) Replace & by and.
7) Check whether there are no quotation marks within arguments for batch commands (e.g. "Volume "1", second quotation mark needs removal). If there are such cases, fix them using the "Replace" function.

The result should look similar to Figure 33. The remaining quotation marks are part of the normal syntax for EXIFTool. Now save the file.

Then go to the "File"-menu, and choose "Rename…". Change the extension to .bat. Click "Save". Notepad++ will recognize the file as a batch file now and add some pretty colours. You can leave the file open in Notepad++ for now.

Preparation of the image files for batch processing

Using Windows Explorer, copy (**not** move) the image files to which you want to add metadata from their original folders to the base folder for batch files.

Figure 34 now shows what an example Batch file folder looks like.

Important: As already mentioned, EXIFTool should be present in the folder from which the batch process will be run.

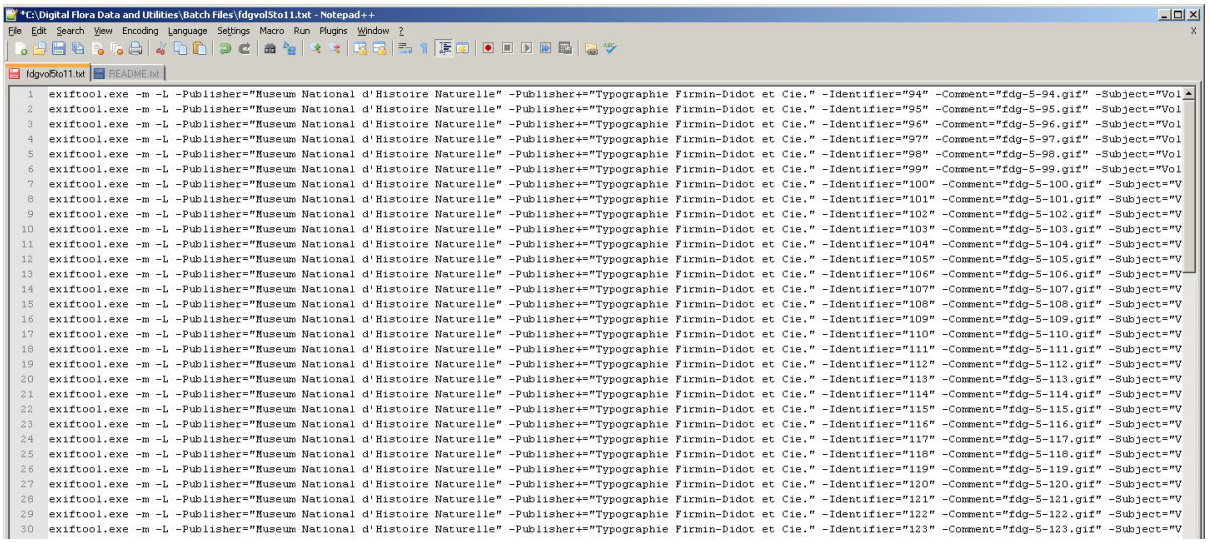


Figure 33: Part of the future batch file with all of the offending quotation marks removed.

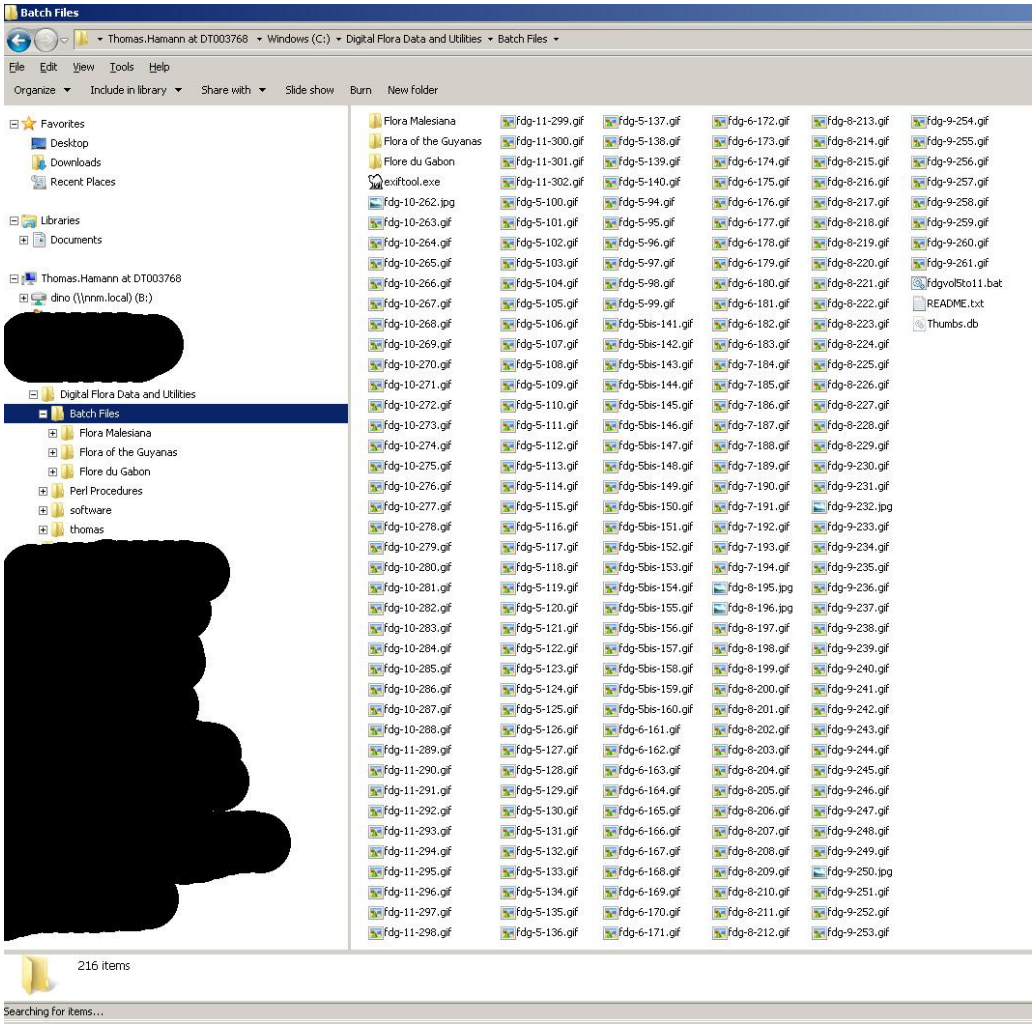


Figure 34: Batch file folder with EXIFTool, image files to which metadata will be added and batch file.

Running the batch process

Double-click the batch file (.bat extension) to run the batch process.

A Windows Command Prompt window (Figure 35) will show up showing you the progress of the batch process, with EXIFTool reporting whether it managed to successfully update the image files and reporting any errors encountered. Keep an eye on this window, as it might give you a hint whether all files are properly updated. If updating a file fails, the first thing to check is whether the file has the same name as what is indicated in the Image Administration and therefore the batch file - the most encountered error was giving the file the wrong extension (.gif or .jpg) in the Image Administration. When the batch process has completed the Windows Command Prompt window will close automatically.

Tip: If you think you saw a file was not updated properly, but it scrolled off the Windows Command Prompt window already, you can hit the "Pause/Break" key on your keyboard (upper right, next to "Scroll Lock") to pause the batch process. Then you can scroll the Windows Command Prompt window and check which file failed. The batch process can be resumed by hitting "Enter". This will not work if the Windows Command Prompt window has closed already.

EXIFTool will save the original file under a new name, consisting of the original file name with _original appended. The updated files will have the same names as the original files. E.g. the original file "*fdg-5-105.gif*" shown in Figure 35 below will be renamed "*fdg-5-105.gif_original*", while the updated file will be called "*fdg-5-105.gif*" again.

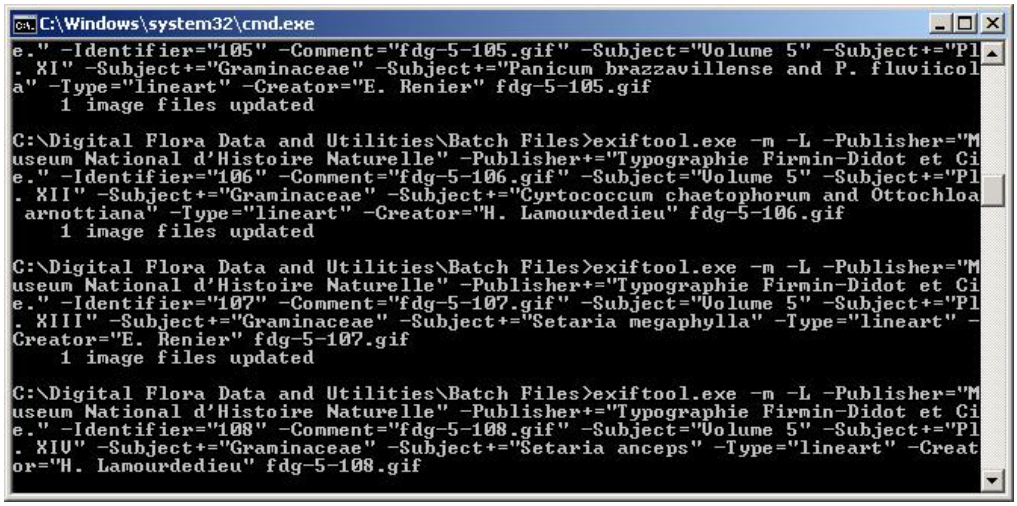


Figure 35: Batch process running.

After running the batch process you can move the files to their respective folders for batch processed files.

Packing for file transfer

To transfer the files to the people who run the CDM server (currently Botanical Garden Botanical Museum Berlin), the files are archived into a ZIP file using either WinZip or Windows' build in compression utility. Each volume of a flora gets one ZIP file containing all images belonging to that volume.

Linking the image files to the treatment files

The image files for a legacy taxonomic treatment have to be linked to from within the XML file (marked up text files) containing the legacy taxonomic treatment text. As mentioned while explaining how the Image Administration was set up, each image file within a Flora has an unique identifier that is repeated in the file name and in the metadata inserted into each image file.

These same identifiers are used in the XML files. Within a XML file, they serve two purposes:

- Linking references to images within the text to the image captions/legends.
- Linking the image captions/legends to the actual image files.

As additional safeguards, the image type and file name (the latter is optional) are also used within the XML mark up for image legends. Figure 36 shows this schematically.

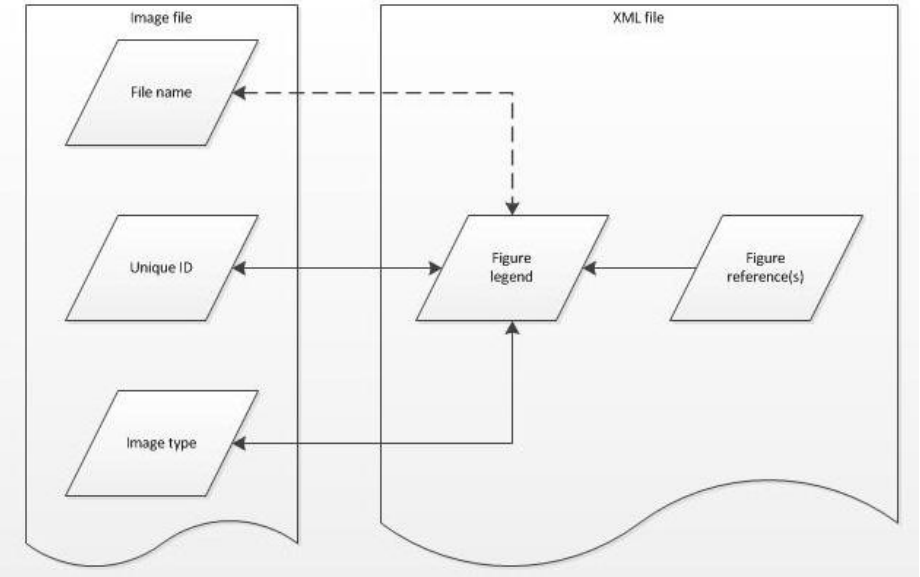


Figure 36: How each image file is linked with the treatment's XML file.

The actual implementation in the XML file works as follows:

- Firstly, each XML file has a URL in its metadata section that refers to the base location for the associated media files (Figure 37). Ask your CDM host what the correct defaultMediaURL has to be.

```
<metaData>
  <defaultMediaUrl>http://wp5.e-taxonomy.eu/media/flora-malesiana/</defaultMediaUrl>
</metaData>
<treatment>
```

Figure 37: The base location of media files is indicated in the Metadata section of an XML file.

- Each figure is usually accompanied by a caption or a legend, a piece of text explaining what can be seen on the figure. This can be used as a placeholder for the actual image in the treatment text. It uses the unique image identifier and also gives the type of image (Figure 38). The filename of the image can optionally be added using an attribute called “url”. This also makes it possible to refer to images located off-site.

```
</feature>
<figure id="ID_3514" type="lineart">Map <num>1</num>. <figureLegend>Distribut
<figure id="ID_3515" type="photo">Plate <num>1</num>. <figureLegend>Pollen o
m>5535</fieldNum></gathering>): h. Alstonia angustiloba Mig. 2-colporate grain (
```

Figure 38: Figure legend in use as placeholder for actual image.

- Each reference to a figure is indicated using a separate XML element that repeats the unique ID in its attribute (Figure 39).

```
scolpium centres. More or less prominent extracolpal nodes occur in Aspidosperma, Geissospermum a
  <figureRef ref="ID_3515">Plate <num>1</num><figurePart>a-c</figurePart>.</figureRef><br />
Vinceae (8 genera: Malouia: Kopsia, Ochrosia, Pleurolfia: Huang 1986, Endress et al. 1996, 2007). Del
```

Figure 39: Reference to a plate.

List of all figures