

Paperwork manual

WORK IN PROGRESS

Contents

1 Settings	3
1.1 Accessing the settings	3
1.2 Work directory	3
1.3 Scanner	3
1.3.1 Default scan source	3
1.3.2 Scanner calibration	3
1.3.3 Scan resolution	3
2 Scanning	4
2.1 Single scan	4
2.2 From feeder	4
2.3 OCR languages	4
2.3.1 Windows	4
2.3.2 Debian	5
2.3.3 Fedora	5
2.3.4 Ubuntu	5
2.4 OCR enabled / disabled	5
3 Importing	5
3.1 Images	5
3.2 PDF	5
3.3 Many PDFs in one shot	6
4 Labels	7
4.1 Creating new labels	7
4.2 Setting labels on documents	7
4.3 Modifying a label	7
4.4 Deleting a label	7
5 Searching	7
5.1 Simple search	7
5.2 Advanced search	7
6 Viewing	7
6.1 View pages as grid	7
6.2 View pages as list	7
6.3 Zoom level	7
7 Exporting	7
7.1 Document	7
7.2 Page	7
8 Printing	7
9 Copying text	7
10 Editing pages	7

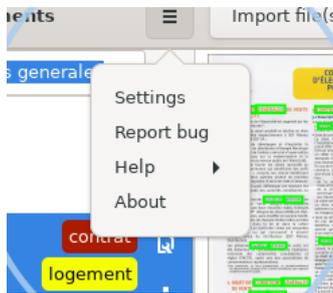
11 Moving pages	7
11.1 inside a document	7
11.2 from a document to another	7
12 Switching to another work directory	7
13 Backup	8
14 Synchronisation	8
14.1 USB key / USB drive	8
14.2 File Synchronization applications	8
14.2.1 DropBox	8
14.2.2 Shared folder	9
15 Encryption	9
15.1 Windows	9
15.2 GNU/Linux	9
15.2.1 Ecryptfs	9
15.2.2 Encfs	9
16 Advanced use and information	10
16.1 Redo OCR	10
16.1.1 On all the documents	10
16.1.2 On one document	10
16.2 Highlight all words	10
16.3 Keyboard shortcuts	10
16.4 Paperwork's files locations	10
16.5 Work directory layout	10
16.5.1 Global organisation	10
16.5.2 hOCR files	12
16.5.3 Label files	12
16.6 Statistics	12
17 Getting support / reporting issues	12
17.1 Diagnostic dialog	12
17.2 Gnome's Gitlab issue tracker	12
17.3 Mailing-list	12
18 Uninstalling	13
18.1 Windows	13
18.2 GNU/Linux	13

1 Settings

1.1 Accessing the settings



Note that, on GNU/Linux, the application menu may be at the top of the screen.



1.2 Work directory

1.3 Scanner

1.3.1 Default scan source

Some scanners have many sources/input. Basic scanners have only one source : Flatbed. Some others have a feeder, allowing them to scan many pages at one.

The settings here is the source to use for single scan.

If you select Flatbed here and use the multi-scan dialog, Paperwork will automatically switch to the feeder (if one is found ; otherwise the default source is used too).

1.3.2 Scanner calibration

Scanners tend to provide images actually bigger than the scanned pages. Since most of the time, you will always scan pages having the same size (A4 or Letter usually), Paperwork provides an option called scanner calibration. Scanner calibration in Paperwork is simply a pre-cropping of the images coming from the scanner.

1.3.3 Scan resolution

Scanner resolution defines how detailed the images coming from your scanner must be.

Higher resolutions mean

- longer scans,
- longer OCR,

- more time to display,
- more space used on disk,
- but also better OCR.

Lower resolutions mean

- shorter scans,
- shorter OCR,
- less time to display,
- less space used on disk,
- but also inferior OCR,
- and possibly unreadable image (even by a human).

300 dpi is considered a good trade-off. You may want to reduce it to 200 dpi on slow computers.

2 Scanning

2.1 Single scan

Pages are appended to the current document.

2.2 From feeder

The option scan from feeder is enabled only if Paperwork has detected a feeder on your scanner.

You have to tell Paperwork how many pages go in each document. If you just want Paperwork to scan pages until none are left, you can just specify a huge number of pages (99 for example).

2.3 OCR languages

By default, Paperwork uses Tesseract for the OCR. If unavailable, it falls back on Cuneiform. On Windows, Tesseract is provided with Paperwork.

To get better results, OCR tool need to know the language used in the document(s).

The language available in the settings dialog of Paperwork are those understood by the OCR tool. If your language is not in the list, it means the OCR tool doesn't have the data required to read your language and you must install them.

2.3.1 Windows

Tesseract and all its data files are provided by Paperwork's installer. If a language is not available in the installer, it either means it hasn't been packaged (in which case you can request it), or there is no data file available yet for this language.

2.3.2 Debian

```
# OCR (Tesseract)
$ sudo apt-get install tesseract-ocr tesseract-ocr-<lang>
```

2.3.3 Fedora

```
# OCR (Tesseract)
$ sudo dnf install tesseract tesseract-langpack-<lang>
```

2.3.4 Ubuntu

```
# OCR (Tesseract)
$ sudo apt-get install tesseract-ocr tesseract-ocr-<lang>
```

2.4 OCR enabled / disabled

When you scan a page using Paperwork, Paperwork will immediately run the OCR on it. This process may take a while for each page.

In case you want to scan a lot of pages quickly (for instance, the first time you use Paperwork), OCR can be temporarily disabled.

OCR can then be run on all the documents managed by Paperwork in one shot.

3 Importing

3.1 Images

Paperwork supports a lot of file formats. It supports JPEG, PNG, GIF, BMP, TIFF, etc.

Each image is considered as a page. Currently, you can only import one file at a time.

Images are always appended to the document currently opened. Simply select an empty document ("New document") to create a new document while importing.

OCR is always run on imported images. If the imported image is the first page of a new document, Paperwork will automatically apply documents labels.

Note that Paperwork is a document manager. While it can, it is not designed to handle images with only very little text or photos. Automatic labeling will not work correctly on such documents.

The OCR (Tesseract) works very well with black text on white background. Automatic labeling uses recognized text and requires as many keywords on the first page as possible.

3.2 PDF

Each PDF is always considered as a whole document. They are never appended to existing document. They are copied as is in the work directory and are never modified by Paperwork (just moved and renamed).

Paperwork will look for pages with no text attached. On those pages, it will automatically run OCR. Once all the pages have been examined, it will

automatically apply document labels. Note that this process may take a few minutes for big PDFs files.

If the PDF is already part of your documents, Paperwork will simply ignore it.

3.3 Many PDFs in one shot

If you import a folder, Paperwork will browse this folder and look for PDFs to import. Already-imported PDFs are simply ignored. Folder is browsed recursively (all the folders inside the folder are also examined).

4 Labels

4.1 Creating new labels

4.2 Setting labels on documents

4.3 Modifying a label

4.4 Deleting a label

5 Searching

5.1 Simple search

5.2 Advanced search

6 Viewing

6.1 View pages as grid

6.2 View pages as list

6.3 Zoom level

7 Exporting

7.1 Document

7.2 Page

8 Printing

9 Copying text

10 Editing pages

11 Moving pages

11.1 inside a document

11.2 from a document to another

12 Switching to another work directory

Before copying or moving the work directory of Paperwork, please close Paperwork.

When Paperwork starts, one of the first things it does is to look for any change in its current work directory. Therefore, if you moved your work directory, when you will restart Paperwork, since it won't find anything, the document list will be empty.

You must then go in the settings and change the work directory location to the new one.

In the following example, we are switching to a work directory contained in a DropBox's folder :

Note that, on GNU/Linux, the application menu may be at the top of the screen.

Paperwork will automatically scan the newly selected work directory, and update its index according to its content.

13 Backup

14 Synchronisation

While Paperwork is a personal document manager, it is not a file synchronization application. However, it is designed to be used with file synchronization applications (Dropbox, OneDrive, Owncloud, SparkleShare, etc).

When you start Paperwork, one of the first things it does is check the content of the work directory. It looks for any changes and updates its document list and index accordingly, automatically.

14.1 USB key / USB drive

This is the simplest way to share documents. Simply copy your work directory to an USB key, tell Paperwork to use it, and you're done.

Beware: You should backup your USB key from time to time on another one.

14.2 File Synchronization applications

Those applications synchronize a local directory with a remote server (or cloud). All the changes you do in your folder are applied on the server. All the changes applied on the servers are applied to the computers that connect to it. The server can belong to you or to someone else (usually a company).

Beware: If you choose to host your documents on someone else server (DropBox, OneDrive, etc), they can access all your documents. Paperwork does not cipher them.

Paperwork is tested daily with SparkleShare. While this is not the easiest one to use, SparkleShare let you host your files yourself. Using DropBox or OneDrive can make sense if you're sharing not-so-confidential documents with others (associations, etc).

14.2.1 DropBox

Here we are detailing the process to use DropBox, but it is similar for other file synchronization applications.

First, you must copy or move your work directory inside the DropBox folder (please stop Paperwork before):

Then you must tell Paperwork to use this new work directory.

14.2.2 Shared folder

If all your computers are on the same network, you can share your work directory. However, be really careful regarding permissions. Being too permissive could let a pirate access all your personal documents ! And setting them correctly is tricky.

Beware: While file synchronization applications usually maintain an historic, shared folders do not. You should do backups from time to time.

Here are the instructions for Microsoft Windows:

On the client side, you must map the shared folder to a drive.

15 Encryption

15.1 Windows

TODO

15.2 GNU/Linux

GNU/Linux distributions include many tools to encrypt whole directories.

With Paperwork, there are 2 directories that should be encrypted to protect your privacy:

- Your work directory (by default `~/papers`, can be changed in the settings)
- The cache directory (`~/.local/share/paperwork`, cannot be changed) (it contains index files from which the content of your documents could be partially recovered)

Note that if you want to be sure that your data are always encrypted, it's recommended to encrypt your whole home directory or even your whole system if possible.

15.2.1 Ecryptfs

On GNU/Linux Debian and Ubuntu, you can easily create a directory `Private` in your home directory. This directory will be encrypted using the password you use to connect when you start your computer. Just type `ecryptfs-setup-private` in a terminal to create it. You have to logout/login again. You can then put the work directory of Paperwork in it.

Once the directory has been created, you can also store Paperwork index in it:

```
$ mv ~/.local/share/paperwork ~/Private/paperwork_index
$ ln -s ~/Private/paperwork_index ~/.local/share/paperwork
```

15.2.2 Encfs

Encfs can also be used to create encrypted directories easily. However, beware that Encfs seems to have some security weaknesses.

```
$ encfs ~/.papers ~/papers
```

16 Advanced use and information

16.1 Redo OCR

16.1.1 On all the documents

16.1.2 On one document

16.2 Highlight all words

16.3 Keyboard shortcuts

- Ctrl+E
- Ctrl+N
- PageUp
- PageDown
- Ctrl+PageUp
- Ctrl+PageDown
- Shift+MouseButton on a document

16.4 Paperwork's files locations

By default:

- Configuration : `~/config/paperwork.conf`
- Index : `~/local/share/paperwork`
- Documents : `~/papers`

(same paths are used on Windows ; `~` = `C:\Users[login]` ; folders are hidden)

The index is always updated according based on the documents in the work directory. When Paperwork starts, the modification time of each file is used to detect changes on the documents.

16.5 Work directory layout

`workdir—rootdir` = `~/papers` (by default)

16.5.1 Global organisation

In the work directory, you have folders, one per document.

The folder names are (usually) the scan/import date of the document: `YYYYM-MDD_hhmm_ss[.idx]`. The suffix `'idx'` is optional and is just a number added in case of name collision.

In every folder you have:

- For image documents:
 - `paper.iXl.jpg` : A page in JPG format (X starts at 1)

- paper.iX_i.words (optional) : A hOCR file, containing all the words found on the page using the OCR (optional, but required for indexing ; can be regenerated with the options "Redo OCR").
 - paper.iX_i.thumb.jpg (optional, generated automatically) : A thumbnail version of the page (faster to load)
 - labels (optional) : a text file containing the labels applied on this document
 - extra.txt (optional) : extra keywords added by the user
- For PDF documents:
 - doc.pdf : the document labels (optional) : a text file containing the labels applied on this document
 - extra.txt (optional) : extra keywords added by the user
 - paper.iX_i.words (optional) : A hOCR file, containing all the words found on the page using the OCR. Some PDF contains crap instead of the real text, so running the OCR on them can sometimes be useful.

Here is an example a work directory organisation:

```
$ find ~/papers
/home/jflesch/papers
/home/jflesch/papers/20130505_1518_00
/home/jflesch/papers/20130505_1518_00/paper.1.jpg
/home/jflesch/papers/20130505_1518_00/paper.1.thumb.jpg
/home/jflesch/papers/20130505_1518_00/paper.1.words
/home/jflesch/papers/20130505_1518_00/paper.2.jpg
/home/jflesch/papers/20130505_1518_00/paper.2.thumb.jpg
/home/jflesch/papers/20130505_1518_00/paper.2.words
/home/jflesch/papers/20130505_1518_00/paper.3.jpg
/home/jflesch/papers/20130505_1518_00/paper.3.thumb.jpg
/home/jflesch/papers/20130505_1518_00/paper.3.words
/home/jflesch/papers/20130505_1518_00/labels
/home/jflesch/papers/20110726_0000_01f
/home/jflesch/papers/20110726_0000_01/paper.1.jpg
/home/jflesch/papers/20110726_0000_01/paper.1.thumb.jpg
/home/jflesch/papers/20110726_0000_01/paper.1.words
/home/jflesch/papers/20110726_0000_01/paper.2.jpg
/home/jflesch/papers/20110726_0000_01/paper.2.thumb.jpg
/home/jflesch/papers/20110726_0000_01/paper.2.words
/home/jflesch/papers/20110726_0000_01/extra.txt
/home/jflesch/papers/20130106_1309_44
/home/jflesch/papers/20130106_1309_44/doc.pdf
/home/jflesch/papers/20130106_1309_44/paper.1.words
/home/jflesch/papers/20130106_1309_44/paper.2.words
/home/jflesch/papers/20130106_1309_44/labels
/home/jflesch/papers/20130106_1309_44/extra.txt
```

16.5.2 hOCR files

With Tesseract, the hOCR file can be obtained with following command:

```
tesseract paper.<X>.jpg paper.<X> -l <lang> hocr && mv paper.<X>.html paper.<X>.words
```

For example:

```
tesseract paper.1.jpg paper.1 -l fra hocr && mv paper.1.html paper.1.words
```

16.5.3 Label files

Here is an example of content of a label file:

```
facture,#0000b1588c61 logement,#f6b6ffff0000
```

It's always [label],[color]. For a same label, the color should always be the same.

16.6 Statistics

You can get various statistics regarding your documents. Just have a look at the diagnostic output. Statistics are close to the end of the output.

17 Getting support / reporting issues

17.1 Diagnostic dialog

When querying support about bugs or lack of scanner support, we may ask you the diagnostic output. This output is basically a lot of informations regarding what happened when you used Paperwork, your scanner(s) and various usage statistics.

Before trying to get this output, you must reproduce the issue you're having. For instance, if you have an error when scanning, you must try scanning before getting the diagnostic output. Do not restart Paperwork between both operations (output would be reset back to zero).

To get this output:

You can then save the output to a file, and send this file to support (either on a ticket you opened, or by email).

17.2 Gnome's Gitlab issue tracker

For bugs and feature requests: <https://gitlab.gnome.org/World/OpenPaperwork/paperwork/issues>

17.3 Mailing-list

For general discussions: <https://gitlab.gnome.org/World/OpenPaperwork/paperwork/wikis/Contact>

Be careful however: By default, Google groups set your subscription to "no email"¹.

¹<https://productforums.google.com/forum/#!topic/apps/3OUIPmzKCi8>

18 Uninstalling

Paperwork can be uninstalled. Uninstalling Paperwork *won't* remove your work directory or documents.

18.1 Windows

18.2 GNU/Linux

If you installed Paperwork manually:

```
sudo pip uninstall paperwork  
sudo pip uninstall pyocr
```

(it's python-pip on some systems)

If you installed many versions of these packages, you may have to run these commands many times.

Note that there are other dependencies installed with Paperwork. However, python-pip can't detect and remove automatically unused dependencies. This is why you should use your distribution package(s) if possible.