Manuel de Paperwork

Table des matières

1   Introduction

La plupart des documents personnels sont plutôt récurrents : fiches de paies, quittances de loyer, facture d'électricité, etc. Pour la plupart des gens mal-organisés, devoir les retrouver plus tard est, au mieux, inquiétant. Pour la plupart des gens organisés, les nommer et les trier est aussi ennuyeux que regarder la pluie tomber.

L'idée principale derrière Paperwork est que gérer les documents est un travail d'ordinateurs. Les humains devraient faire aussi peu que possible pendant que les machines font le gros du travail. L'objectif final ici est "scanner & oublier".

Si vous cherchez un logiciel qui vous laissera nommer les documents individuellement, les organiser dans des hiérarchies complexes, les marquer manuellement à chaque fois, corriger les ratés mineurs de l'OCR, etc, alors Paperwork n'est pas pour vous.

# 2   Définitions

2.1   Répertoire de travail

Paperwork enregistre tous vos documents dans un seul répertoire : Le répertoire de travail. Dans ce répertoire, chaque document a son propre sous-répertoire.

Bien que cela rende Paperwork difficile à utiliser avec d'autres outils, ça a un avantage majeur : Vous n'avez pas à vous inquiéter des noms des fichiers et de la structure des répertoires.

2.2 Document
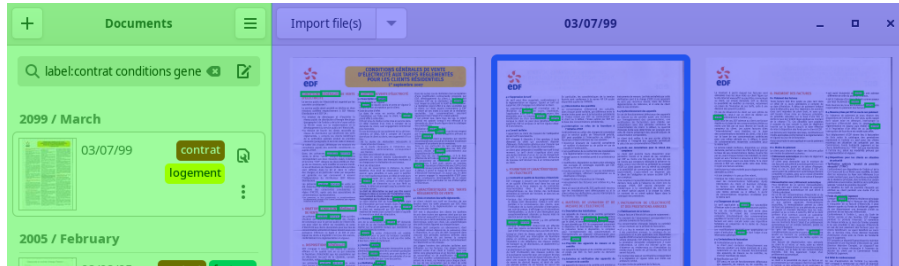
Dans Paperwork, un document est ensemble de pages. Sur le disque, ça peut être un jeu de fichiers JPEG ou un fichier PDF.

Chaque document est identifié uniquement par une date. Il peut s'agir de la date à laquelle vous l'avez importé (par défaut) ou une date de votre choix.

Ils sont affichés sur le coté gauche de la fenêtre principale (partie en vert sur la capture d'écran ci-dessous).

2.3 Page

Dans Paperwork, une page est juste une image et les positions des mots sur cette image.

Les images peuvent venir d'un scanner ou être importées. Dans ces cas, elles sont enregistrées sous forme de fichiers JPEG et leur texte est extrait grâce à la ROC (Reconnaissance Optique de Caractères). ROC est un processus assez long. Il peut prendre jusqu'à plusieurs minutes pour chaque page. Par conséquent, le texte extrait est enregistré dans des fichiers hOCR à coté des fichiers JPEG.

Les pages peuvent aussi être les pages d'un fichier PDF. Dans ce cas, par défaut, Paperwork enregistre simplement une copie du fichier PDF.

Paperwork ne mémorise pas si une page est le recto ou le verso. Paperwork ne mémorise pas la taille du papier (A4, Letter, etc) correspondant à la page.

Les pages sont affichées sur le coté droit de la fenêtre principale (la partie en bleu sur la capture d'écran ci-dessus).

2.4 Indexation et mots-clés
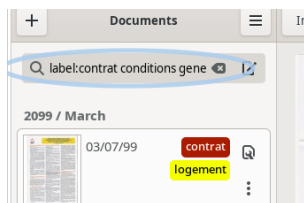
Bien sûr, vous avez besoin de pouvoir retrouver vos documents. Paperwork gère un index avec tout les mots-clés trouvés dans vos documents.

Tapez juste quelques mots-clés, et vous retrouverez vos documents.

2.5 Étiquettes et mot-clés additionnels
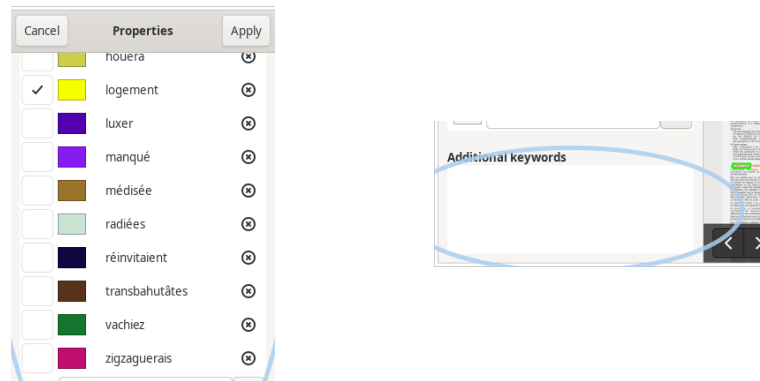
Unfortunately, sometimes, documents don't contain the keywords needed to find them back. Also OCR is not a perfectly realiable process and may not work.

To mitigate those issues, you can add labels (or tags) on your documents and provide additional keywords. Both are added to the index.

Labels are displayed beside documents. Additional keywords are almost never displayed.

3 Réglages

3.1 Accéder aux réglages
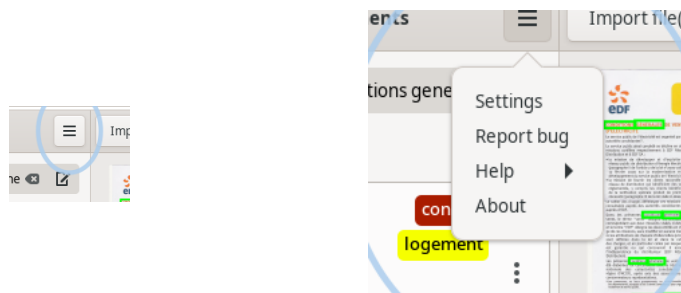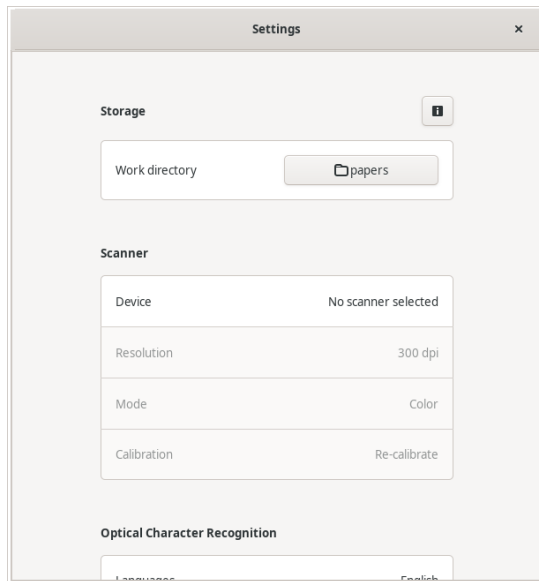
3.2 Répertoire de travail
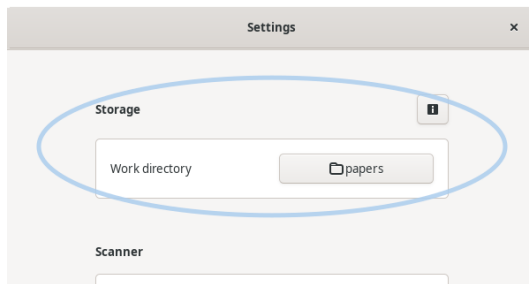
The work directory is the directory where you want all your documents stored. It can be a standard folder, a folder synchronized across multiple computers or on a network share.

Once you close the settings dialog, the work directory will be scanned and Paperwork index will be updated according to its index.

Each time Paperwork starts, it will look for changes in this folder and synchronize its index accordingly.

3.3 Scanner
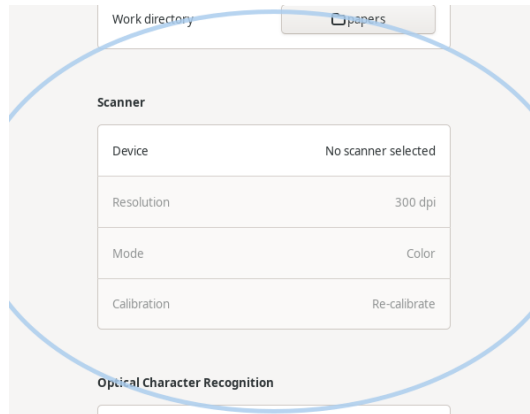
3.3.1 Périphérique

When starting, Paperwork looks for scanners. The scanner to use can be selected in the settings.

Webcams, file storage, etc, cannot be used. Only paper-eaters.

3.3.2 Mode du scanner

Most modern scanners scan in color in a reasonable time. However some older scanners scan much faster in grayscale or even in black&white. Here you can select the mode to use.

3.3.3   Résolution du scan

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.

3.3.4   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 an area that will always be cropped out of images coming from the scanner.
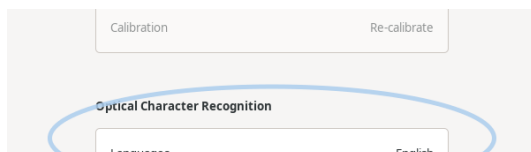
3.4   OCR

By default, Paperwork uses Tesseract for the OCR. If unavailable, it falls back on Cuneiform.

On Linux, if installed with Flatpak, Paperwork is always provided with Tesseract. On Windows, Paperwork is always provided with Tesseract.

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.

3.4.1   Adding languages

Flatpak

```
# <langs> is a list of 2-letters language codes separated ';'
# ex: en;fr;de
flatpak config --user --set languages "<langs>"
flatpak update --user
```

Debian

```
# <lang> is a 3-letter language code
# ex: 'fra' for French
$ sudo apt-get install tesseract-ocr tesseract-ocr-<lang>
```

Fedora

```
# <lang> is a 3-letter language code
# ex: 'fra' for French
$ sudo dnf install tesseract tesseract-langpack-<lang>
```

Ubuntu

```
# <lang> is a 3-letter language code
# ex: 'fra' for French
$ sudo apt-get install tesseract-ocr tesseract-ocr-<lang>
```

Windows   Tesseract and all its data files are provided by Paperwork's installer. You can rerun the installer to install other languages.

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.

3.4.2 Disabling OCR

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. To disable OCR, you simply have to unselect all OCR languages.

3.5 Mises à jour
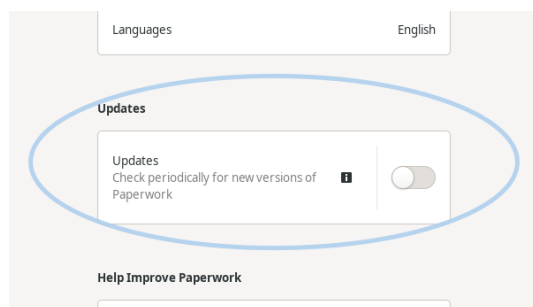
If you enable this option, when Paperwork starts, Paperwork will look for updates if it hasn't done so for a week or more. To know if a new version is available, it has to send an HTTPS query to 'openpaper.work'.

If an update is found, it will notify you but it won't install it.

4   Nouveau document

By default, in the document list, Paperwork includes a document called "New document". If you open it, it always appears empty. This document actually doesn't exist yet on disk, but will exist as soon as you put a page in it. You can add pages in it by scanning, importing file(s) or dropping a page from another in it.

As soon as you put any content in it, this document will get its own date (the current one by default). In the document list, "New document" will be replaced by this date, and a new "New document" will be added to the document list.

If you are currently searching something (see the chapter "Searching"), only search results are displayed and therefore this "New document" isn't displayed.

You can get it back by clicking the button "+" in the top left corner of the main window.

5 Scanning
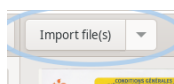
If a scanner has been selected in the settings, you can use it to scan pages.

In the header bar, there is a button to add pages. The small arrow on the right gives access to possible page sources. Those page sources include your scanner sources (Flatbed, Feeder).

Once you've selected the scanner source you want to use, you can click on the button "Scan from ...".

This will start a scan session :

— Scanned pages are appended at the end of the current document. If you use a feeder, Paperwork will scan pages until the feeder is empty.
— Paperwork will then crop them according to scanner calibration.
— Paperwork will run OCR on them
— Paperwork will index them

If this scan session creates a new document, Paperwork will try to set labels automatically on the document.

6 Importing

6.1 Images

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

Each image file is considered as a page.

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.

6.2 PDF

Each PDF is always considered as a whole document. They are never appended to existing document. They are copied and renamed in the work directory, but their content is not modified. Paperwork always keeps the original PDF file as is, even if you edit some of its pages : the edited pages are stored beside the PDF file.

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.

6.3 Many PDFs in one shot

When importing, if you select 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).

7 Labels
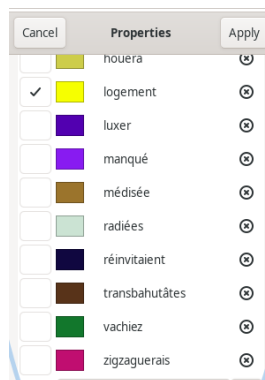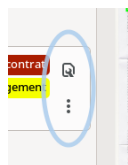
There is currently one constraint in Paperwork : Each label must be on at least one document. Otherwise, when you will restart Paperwork, labels without documents will disappear.

7.1 Creating new labels
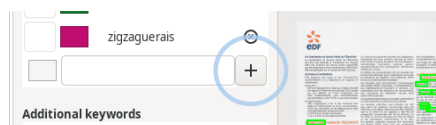
You can click on the gray rectangle on the left side to pick the label color. You can enter the label name in text field between the gray rectangle and the button "+".

Once you click on the button "+", the label will be added to the current document.

The label is actually added once you close document properties. Paperwork will then update its index accordingly.

7.2 Setting labels on documents

When you open document properties, the label list appears. On the left side of each label color, you have a button. This button allows you to add or remove labels on the current document.

The changes are actually written on disk once you close the document properties. Paperwork will update its index accordinly.

7.3 Modifying a label color

When you open document properties, you can click on a label color to change it. A dialog will let you pick the new color.

Label color will actually be changed on disk when you close the document properties. Paperwork will then update the label on all the documents that use it.

7.4 Modifying a label name

When you open document properties, you can click on a label string to change it. A dialog will let you type in the new name.

Label name will actually be changed on disk when you close the document properties. Paperwork will then update the label on all the documents that use it and then reindex them all.

7.5 Deleting a label

To the right of each label is white-on-black cross button. Clicking on it will allow you to delete a label.

Once you will close the document properties, the label will be removed from all the documents having it. Paperwork will then update its index accordingly.

Beware : Once you have closed document properties, there is no way to put back the deleted label.

7.6 Automatic label guessing

Paperwork does use artificial intelligence. It uses a fairly simple method actually : Naive Bayes classifiers. It's the same technology used by email clients to classify mails as spam/non-spam.

Based on all the keywords in all your documents that have (or haven't) a label, it can estimate a probability that a document containing the same keywords should have or shouldn't have this same label. If the probability is high enough, it puts the label on the document automatically when you import it or scan it.

Of course, this approach means that Paperwork needs enough samples to work reliably. You can expect it to start working once you have about 100

documents or more (and only for labels that are on more than 10 documents or more).

8   Searching

8.1   Simple search
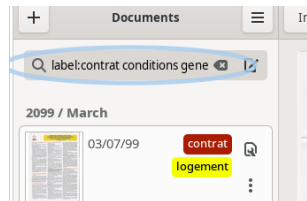
You simply enter keywords in the search field. In a few seconds, you will get all the documents containing those keywords.

Paperwork does a "fuzzy" search : documents with keywords close to the one you gave but not identical are also returned (for instance, 'flech' instead of 'flesch').

You can also use Whoosh query language to make more complex queries. If you want examples, you can use the advanced search dialog described below.

8.2   Recherche avancée
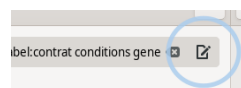
The advanced search dialog helps creating complex search queries. You can specify various criterias and once you click on the apply button, it will generate a search query for you and put it immediately in the search field. Search results will immediately be refreshed as well.

9   Viewing
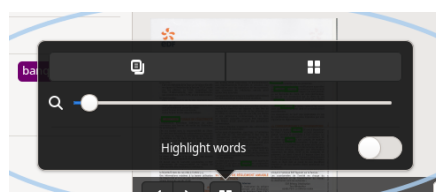
9.1 Zoom level
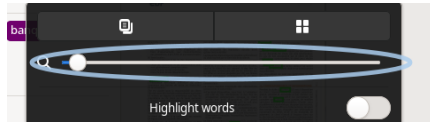
You can change the scale at which pages are displayed using this control.

9.2 View pages as grid

When clicking this button, Paperwork will try to display pages on 3 columns. In this mode, you can drag'n'drop pages to move them inside the document or to another document.

9.3 View pages as list
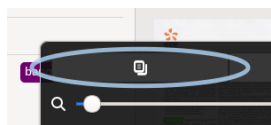
When clicking this button, pages will be scaled so their width is the maximum width allowed by the main window. In this mode, you can select text in the page (and then copy it).

9.4 Highlight all words

This option allows to see quickly all the words identified by OCR. Sometimes (rarely) OCR misses entire chunk in a page. This option allow to see such chunk quickly.

10 Moving pages

10.1 Inside a document

You must display the document pages as a grid (See 9.2).

You can then grab a page (hold the left click button), drag it and drop it wherever you want in a document. While dragging, a blue marker will show you where the page would drop if you release the left click button of your mouse.

10.2   From a document to another

You must display the document pages as a grid (See 9.2).

You can then grab a page (hold the left click button), drag it and drop it in the document list, on the document in which you want the page to go.

11   Copying text

You must display the document pages as a list (See 9.3).

You can then select text in a page. Hold the left click button to start selecting, mouse the mouse cursor to select more words, then release it.

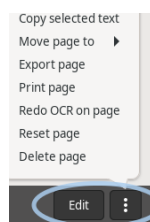You can then copy the selected text, either by pressing Ctrl-C or by using the page menu at the bottom right of the main window. Once copied, you can paste the selected text in any other application (Ctrl-V).

12   Editing a page
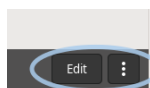
Paperwork includes a very simple image editor. It provides 4 functions :
— Recadrage
— Rotating the page by 90°(can be rotated multiple times)
— Rotating the page by -90°(can be rotated multiple times)
— Automatic Color Equalization : An algorithm that adjust the image brightness, contrast and colors to make it as readable as possible.

13   Reseting a page

Reseting a page returns it to its state when it was scanned or imported, before any pre-processing did occur.

This can be helpful if you made a bad modification on the page (cropped a wrong area for instance), if the calibration settings weren't appropriate or if pre-processing algorithms messed up the page.

14 Effacer

When deleting either documents or pages, they are actually moved in the trash bin of your computer.

Important note regarding Flatpak : A bug may prevent Paperwork from moving files to the trash (we are working on it). In that case, Paperwork will delete the file directly (no recovery possible).

15 Exporting

You can export both documents or single pages.

In both cases, various transformations can be applied before actually exporting them. For instance, you can turn color pages into grayscale pages before putting them in a brand new PDF (making the resulting PDF smaller).

16 Printing

You can print both documents or single pages.

Beware that pages are always sent as images to your printer. So for very big documents, a few minutes may go by before the actual printing start.

17 Backup

18 Synchronisation between multiple computers

While Paperwork is a personal document manager, it is not a file synchronization application. They are applications dedicated to file synchronization that already do that very well. Therefore Paperwork is designed to be used with such applications (Nextcloud, Dropbox, OneDrive, 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. So if another instance of Paperwork on another computer modified something in the work directory and if this change has been synchronized on another computer, the other Paperwork will automatically pick up this change when starting.

18.1 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.

18.2 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 encrypt them.

Paperwork is tested daily with Nextcloud. While this is not the easiest one to install, Nextcloud let you host your files yourself. There are other self-hosted alternatives that exist : SparkleShare, Syncthing, etc.

Using DropBox or OneDrive can make sense if you're sharing not-so-confidential documents with others (associations, etc).

18.2.1 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 : Using a shared folder means having a single copy of your work directory. You should do regular backups of your work directory.

19 Encryption

19.1 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/paperwork2, 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.

19.1.1 cryptsetup

Most GNU/Linux distribution installer now provide an optio4n to encrypt your whole system or your whole /home with cryptsetup . This is the recommended method to protect your documents.

19.1.2 Encfs

Encfs can also be used to create encrypted directories easily.

Beware that Encfs seems to have some security weaknesses. So, while it's probably enough to prevent a laptop thief from accessing your documents, it's likely to be not enough to prevent the NSA or the police from doing so ;-).

```
$ encfs ˜/.local/share/.paperwork2 ˜/.local/share/paperwork2
$ encfs ˜/.papers ˜/papers
```

19.2 Windows

On Windows, you're strongly advised to enable BitLocker to protect your documents. If unavailable, there are other applications (Veracrypt, etc).

20 Keyboard shortcuts
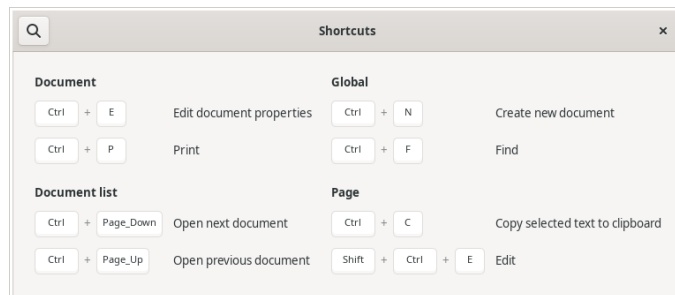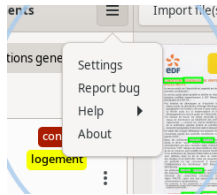
Keyboard shortcuts can be seen by opening the application menu, selecting "Help" and then "Shortcuts".

20.1 Paperwork's files locations

By default :
— Configuration : ˜/.config/paperwork2.conf
— Index : ˜/.local/share/paperwork2
— 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.

20.2 Work directory layout

workdir|rootdir = ˜/papers (by default)

20.2.1 Global organisation

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

The folder names are (usually) the scan/import date of the document : YYYYMMDD_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.<X>.jpg : The original page in JPG format (X starts at 1)
  — paper.<X>.edited.jpg (optional) : The page as edited by the user (X starts at 1)
  — paper.<X>.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.1.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
  — paper.<X>.edited.jpg (optional) : The page as edited by the user (X starts at 1)
  — extra.txt (optional) : extra keywords added by the user
  — paper.<X>.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.
  — passwd.txt (optional) : PDF password, if the PDF is password-protected.
  — doc.docx / doc.odt / ... (optional) : Original file. Converted into PDF (doc.pdf) so Paperwork can parse and display it more quickly.

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.edited.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.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.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.thumb.jpg
```

```
/home/jflesch/papers/20130106_1309_44/paper.2.edited.jpg
/home/jflesch/papers/20130106_1309_44/paper.2.words
/home/jflesch/papers/20130106_1309_44/labels
/home/jflesch/papers/20130106_1309_44/extra.txt
/home/jflesch/papers/20130106_1309_44/passwd.txt
/home/jflesch/papers/20130520_1309_44
/home/jflesch/papers/20130520_1309_44/doc.pdf
/home/jflesch/papers/20130520_1309_44/doc.docx
/home/jflesch/papers/20130520_1309_44/labels
```

20.2.2 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
```

20.2.3 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.

21 Obtenir de l'aide

A forum dedicated to Paperwork exists : https://forum.openpaper.work.
There is also an IRC channel for live discussions : Liberachat, channel #paperwork

If you have questions regarding Paperwork or simply want to chat, those are the places to go.

22 Reporting issues

If you noticed a bug in Paperwork (and you are sure it's a bug), you can make a bug report.

22.1 Bug Tracker

One way to create bug reports is to create tickets on Paperwork bug tracker: https://gitlab.gnome.org/World/OpenPaperwork/paperwork/issues.

This is the recommended way to submit a bug report if you would like to discuss it with Paperwork developpers.

To make sure you include all the required informations, you can use the tool integrated in Paperwork (see below).

22.2   Automatic bug report
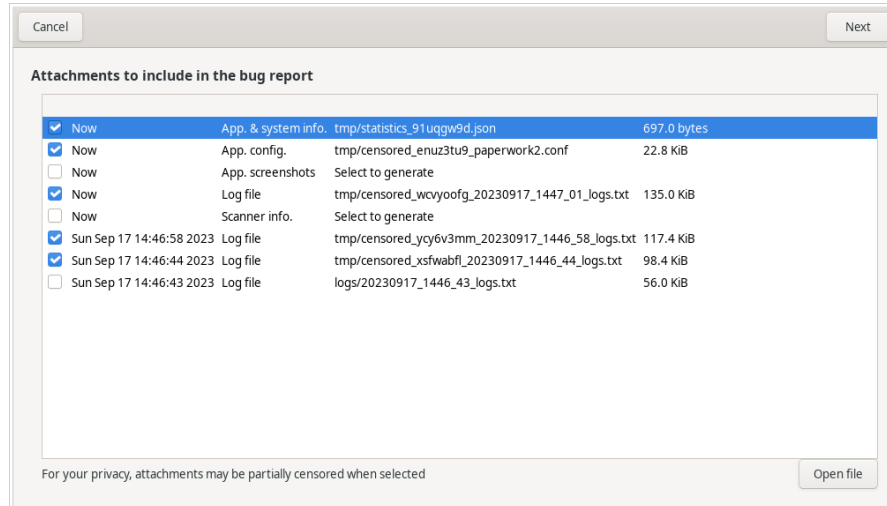
Paperwork includes a tool to make reporting bugs easier. It allows you to get easily all the required information to make a perfect bug report.

All attachments are automatically censored to protect your privacy : Document contents are blurred in screenshots and logs are censored to remove your user name.

If the bug you want to report is related to scanners, please include "Scanner info." in the bug report files.

If te bug you want to report is related to a display problem, please include "App. screenshots" in the bug report files.

22.2.1   Fichier ZIP

You can then obtain a ZIP file with all the data. Please make sure the content of the ZIP file does not contain private information (it shouldn't, but better safe than sorry). Then you can add this ZIP file to a ticket on Gitlab.

22.2.2   Automatic submission

You can also let the tool submit the bug report to openpaper.work automatically. In that case, you won't be able to discuss the bug with developers (or you have to leave a way to contact you in the bug report).

If you use the automatic submission , the tool will give you an URL to see the submitted bug report. This URL is private and shouldn't be shared until you made sure there is no private information in the bug report. If there is private information, you can request deletion of the bug report by sending an email to jflesch@openpaper.work (please specify the private URI in your mail so we can be sure that you are the one who submitted the bug report).

23 Uninstalling

Paperwork can be uninstalled. Uninstalling Paperwork *will never* remove your work directory or your documents.

23.1 GNU/Linux

If you installed Paperwork using the package manager from your distribution (the recommended way), the uninstallation method depends on the package manager.

For instance, on GNU/Linux Debian or GNU/Linux Ubuntu, the following command will take care of it :

```
sudo apt remove --purge paperwork-\*
```

If you installed it using Flatpak, you can use the following command :

```
flatpak --user uninstall work.openpaper.Paperwork
```

23.2 Windows 10

Paperwork can be uninstalled as any Windows applications, by going in Windows Control Panel, clicking on "Applications", finding Paperwork in the list, and then clicking on "uninstall".