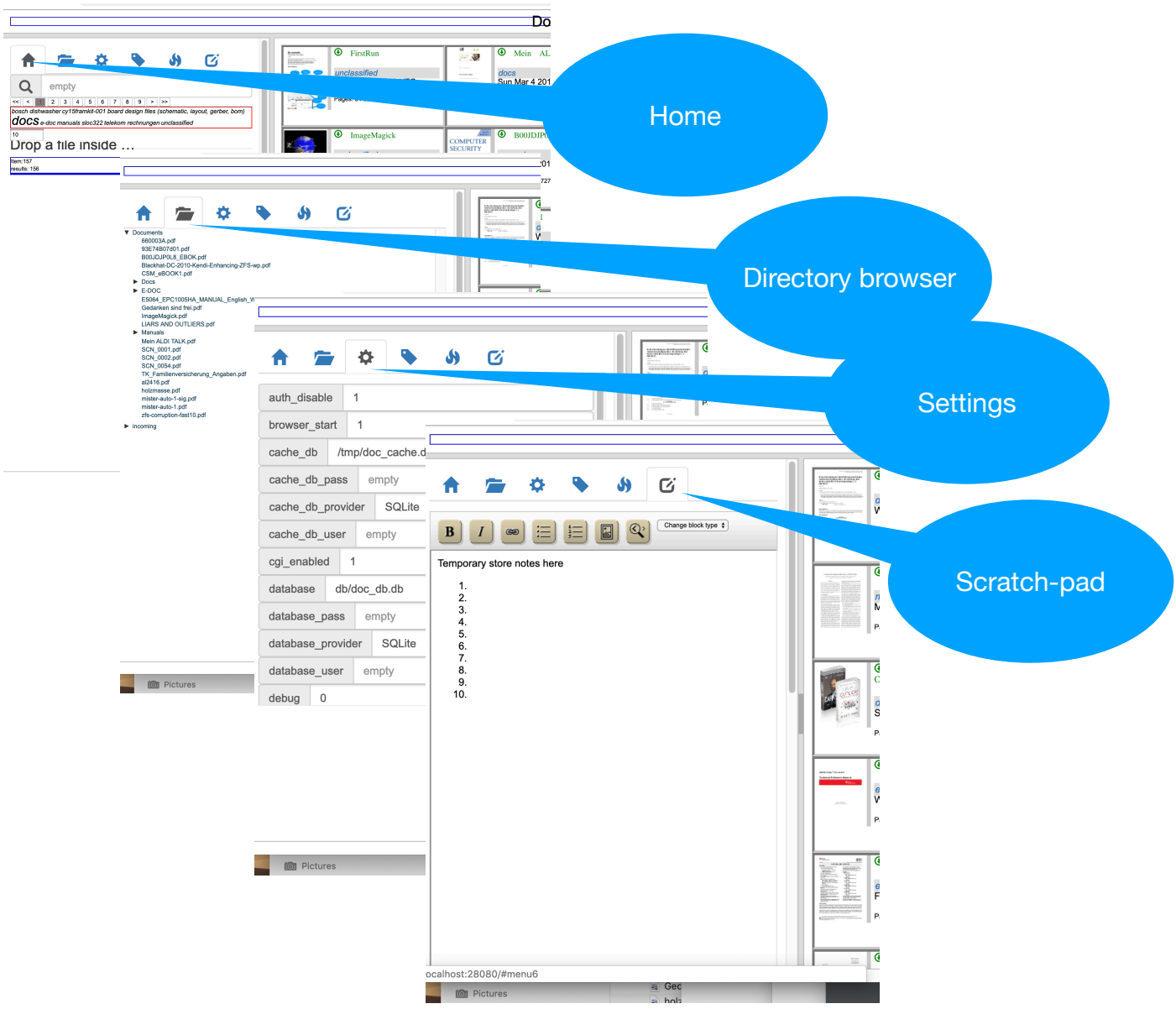
Documentix

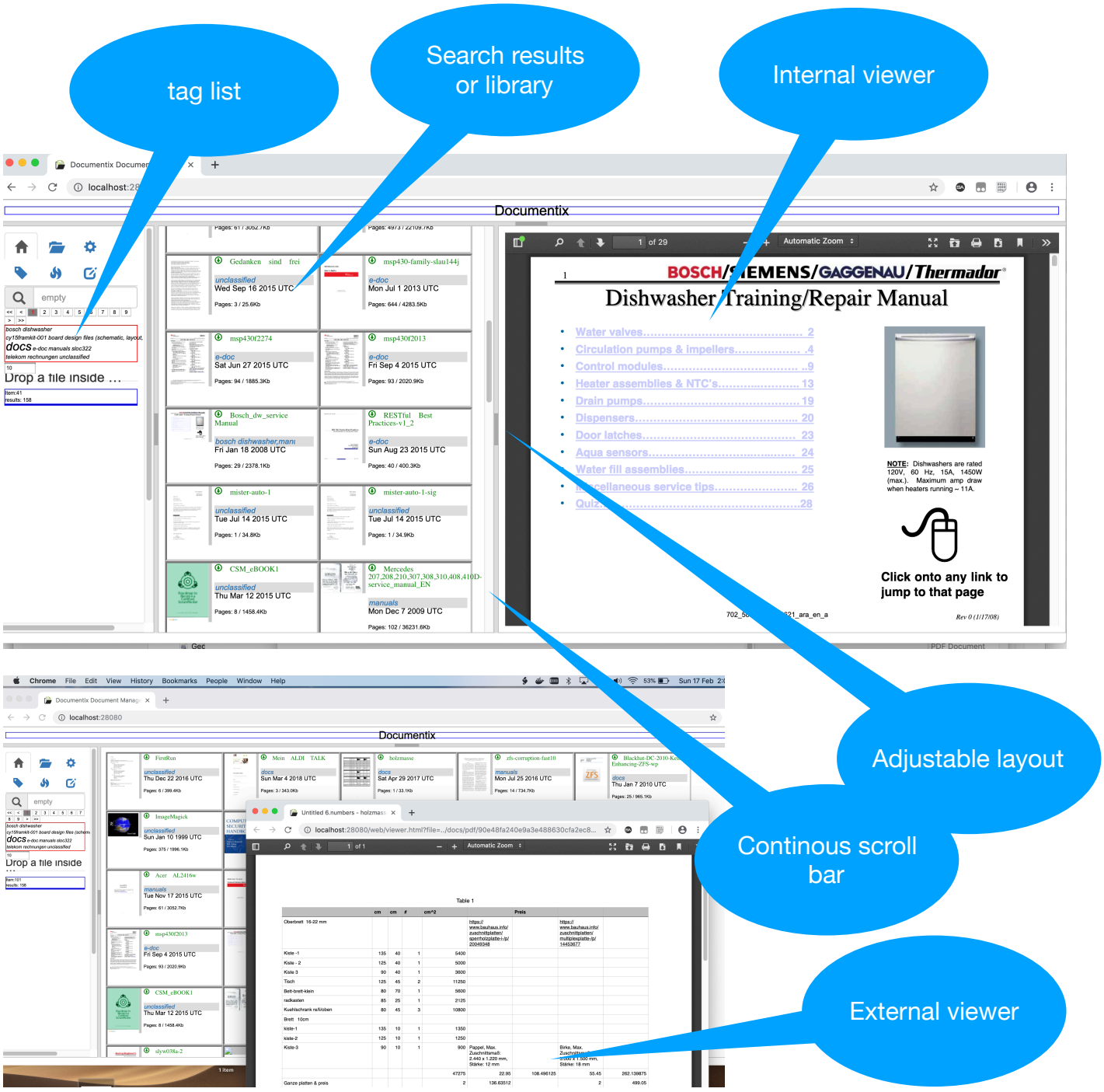
the Document manager

Overview
Quickly searches your collection using a full-text search engine
Lets you classify documents
Automatically classify new documents based on previous training
Automatically OCR's new documents to enable searching
Manages multiple types of documents (pdf/word/ppt/jpeg..)
Presents all archived documents through a web-interface
Allows easy sharing of documents

Tabs available

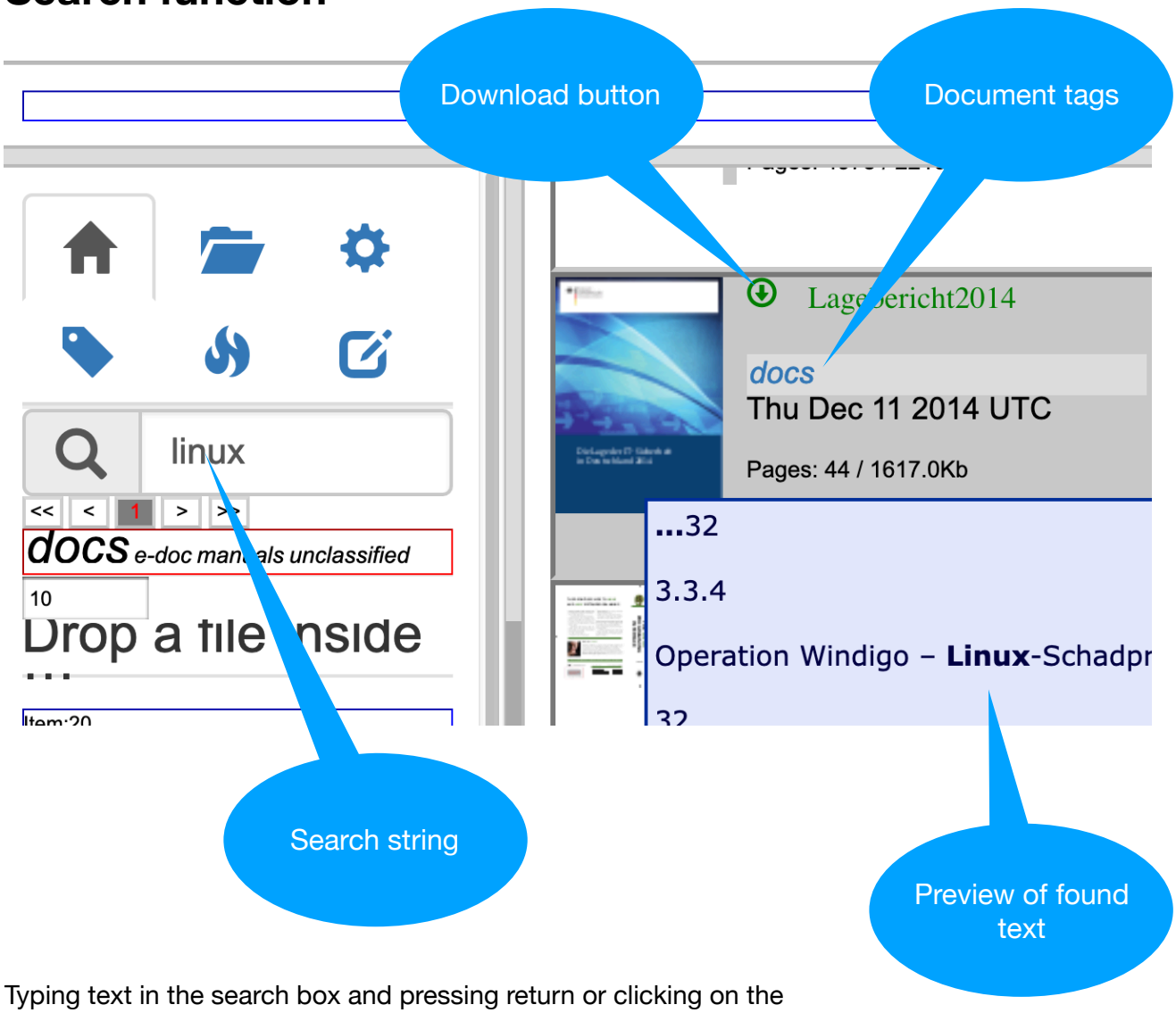


Home tab



The sections can be varies in size to taste.
If the internal viewer is closed, an external tab will be used for viewing.

Search function



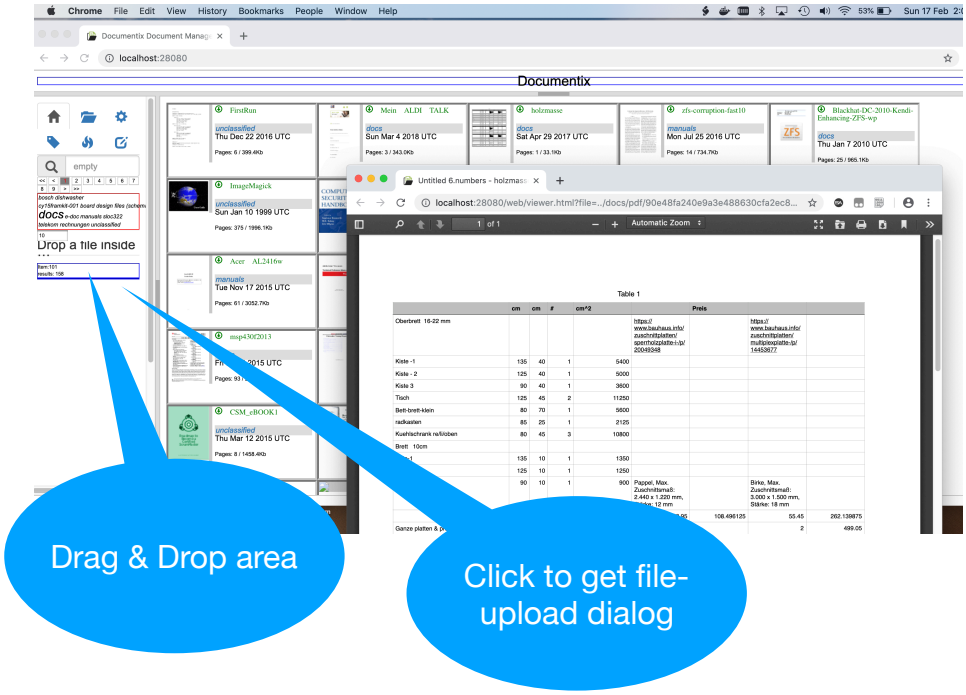
Typing text in the search box and pressing return or clicking on the magnifier will start a full-text search on the documents.

Hovering the mouse over the found items in the search results will show the matching text in the contest of the document.

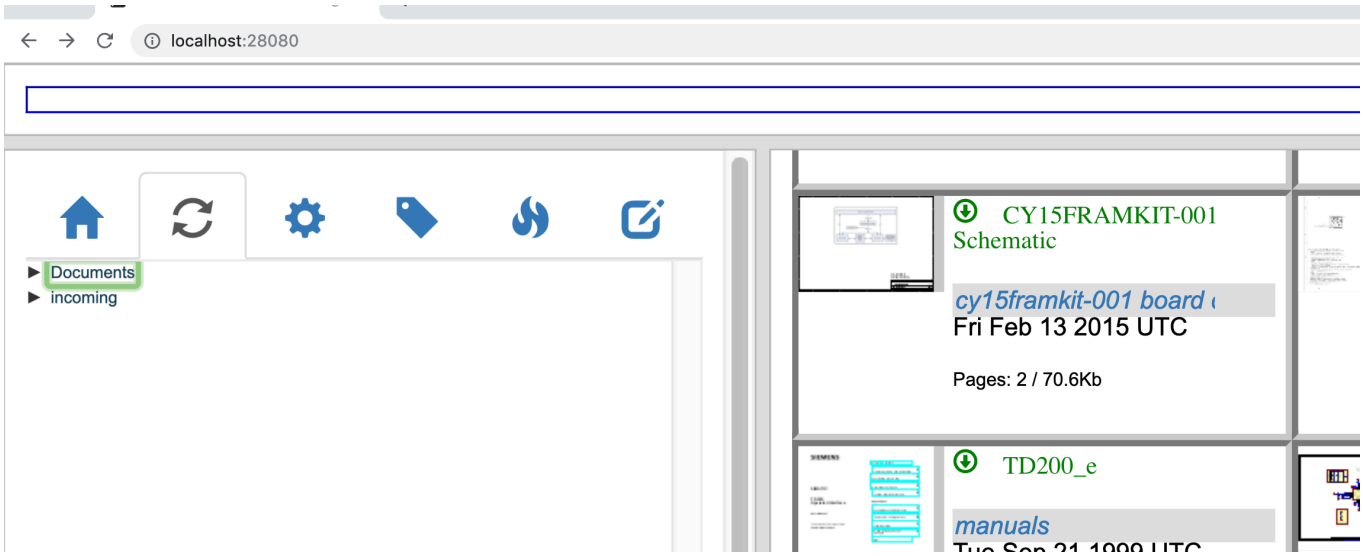
Clicking on the “down-arrow” will download the original document.
Clicking anywhere else will open the searchable document in the viewer.

Clicking on the document tags box, allows creating or deleting new tags.

Upload new documents

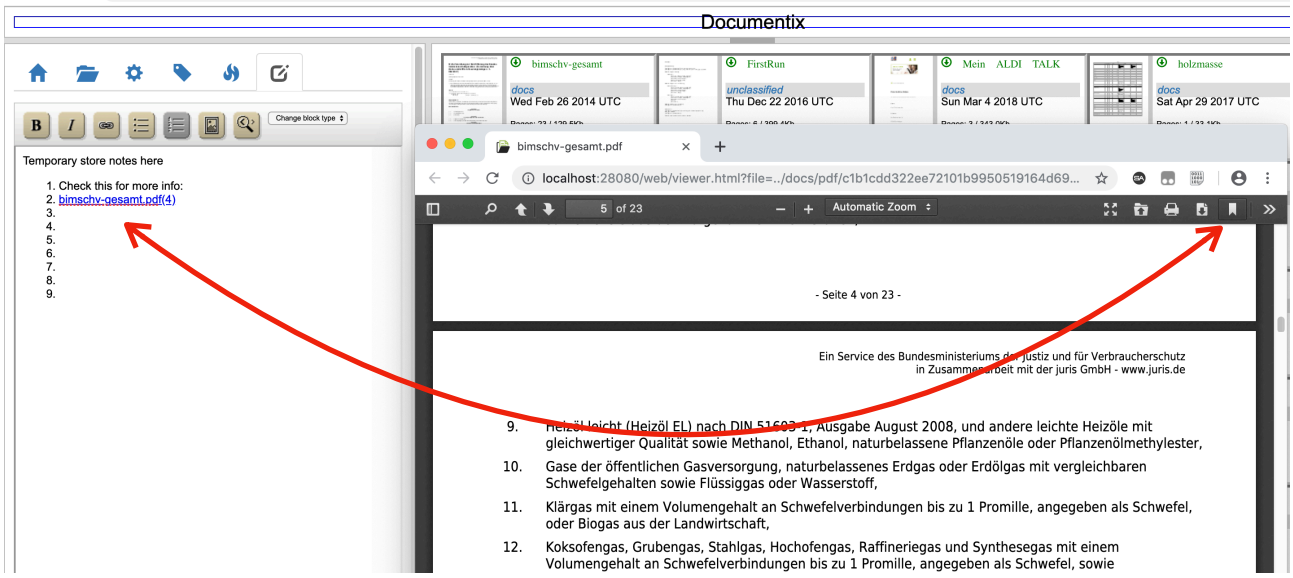


Drag and drop of documents from the finder/explorer uploads these to the server.
Clicking in the box opens upload dialog.



Double clock on a folder in the directory view will start a scan of the below files.
Default tags for files in these folders will be according to the directory structure.
Example:
Documents/Quotes/OEM1/projectA/file1
Will add tags “Quotes”, “OEM1”, “projectA” to the file after scanning.

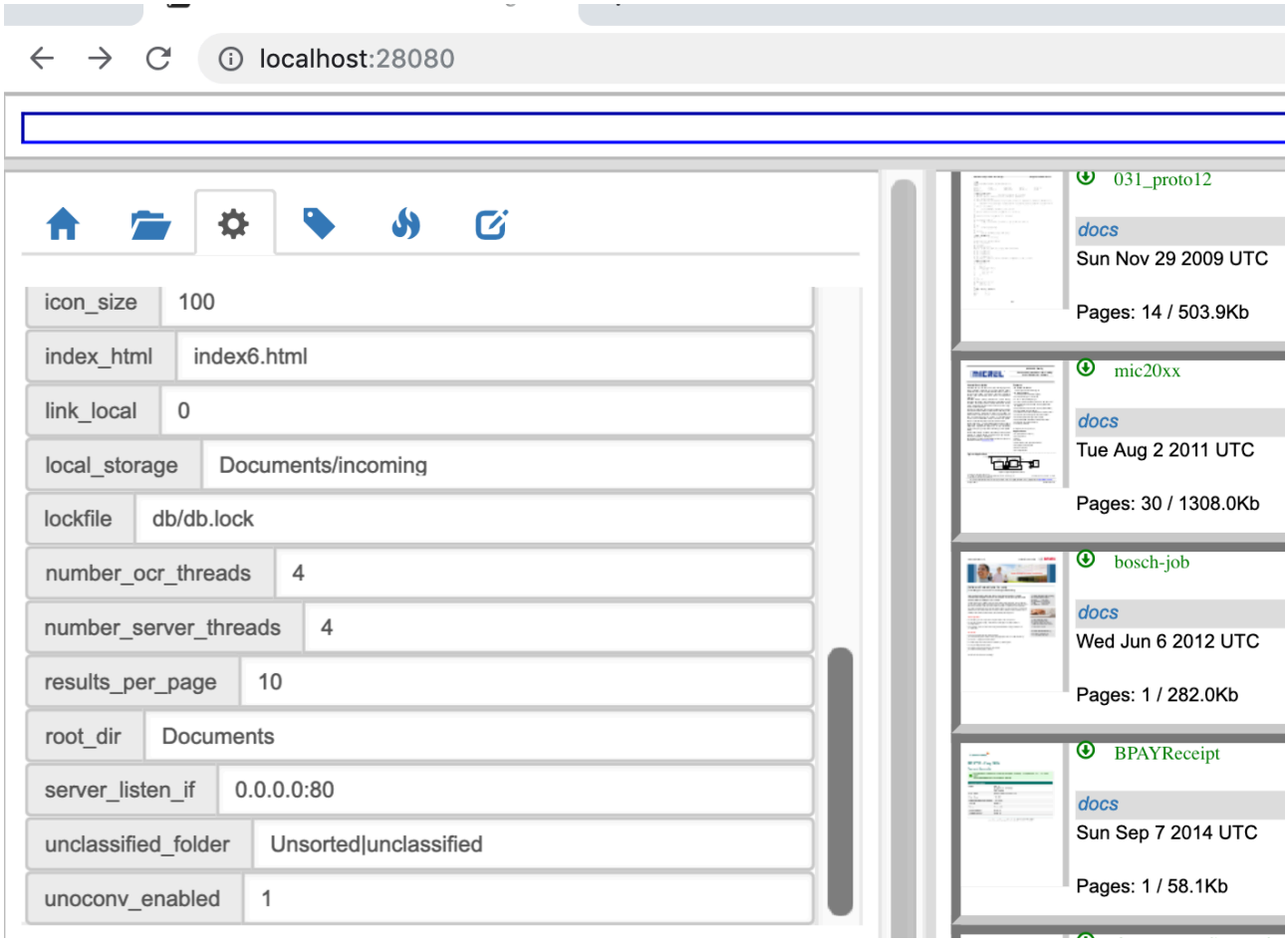
Drag & Drop Bookmark-icon to capture document & location in scratch-pad



The small “bookmark” icon in the viewer allows to capture the exact view in the viewed document. This march can be either dragged & dropped into the “note-pad” (for convenience) or even dragged into excel or other external files. Depending on the application the actual URL is hidden behind the document name and page in brackets.

Some applications (particularly on Windows) only work with the short description if copied from the notepad.

Settings



Currently settings are not very useful unless you know what you are doing. And some of these settings require a restart of the server.

Internals

Sources are available at: https://github.com/thilo-hub/documentix

Documentix uses SQLite and the full-text-search engine as a database backend.
Its primary programming language is perl and A LOT of other wet processing tools.

All documents are stored using the MD5 hash as an index. (Don't tell me it's broken - it does not matter here)

Documents send to the processing engine are converted to PDF files (if not already). If the PDF does not contain more than a certain number of words, tesseract as an OCR engine is started on the document and a searchable PDF is created.

An interface to a Canon-mx870 scanner exists but is not supported in the docker image.

It should be easy to create a curl command-line to upload new scans like:
> curl -F file=@o-page-14.pdf -H x-file-name:MyPage.pdf http://localhost:28080/upload

Which will upload the file "o-page-14.pdf" and name it on the server "MyPage.pdf"

Check the "dockerfile" to understand how the script is started outside of docker.

Authentication is only partly built into the tool, don't rely on it!

If you want to install it on a public facing web-server, use something like haproxy or apache to split the content. The concept is, if you know the MD5 then you can have the file. This is probably ok from a security perspective, since you only know the hash if you have either gotten the hash from someone or you calculated it yourself. In the latter case, you must have the document and in the former someone gave you the complete handle to the document.

This makes all URL's {server&port}/docs/…. possible to be shared on the web.
All other URL's need to be blocked, as search results would leak the MD5.

Retrieving files is possible through the web-interface, if you know the MD5 has of a file.

Lets assume the MD5="009914b6b6a4cb4a37f2fc5bdbc561ca" is.
And the server is reachable under: http://localhost:28080

Then:

http://localhost:28080/docs/raw/009914b6b6a4cb4a37f2fc5bdbc561ca/MyPage
  Will return the original document as it was uploaded

http://localhost:28080/docs/pdf/009914b6b6a4cb4a37f2fc5bdbc561ca/MyPage
  Will get the file as a PDF (converted if necessary including OCR)

http://localhost:28080/web/viewer.html?file=../docs/pdf/009914b6b6a4cb4a37f2fc5bdbc561ca/MyPage
  Will show the file as a PDF in the html viewer.

http://localhost:28080/docs/ico/009914b6b6a4cb4a37f2fc5bdbc561ca/MyPage
  Will return the original document as it was uploaded