OCR4all

A semi-automatic Open Source Software aimed at the OCR of Historical Printings

Installation guide via VirtualBox

Version 1.0, May 2020

In order to stay up to date and informed of future software updates and other technical improvements, we recommend you subscribe to our OCR4all mailing list.

Preliminaries

The present installation guide makes it easy and quick for you to install OCR4all on your device. In order to guarantee the software is running smoothly, OCR4all has been integrated into a so-called virtual machine. The latter runs on your device as its own, little operating system (OS), which is activated every time you start OCR4all.
The following preliminary steps are based on a VirtualBox software download - which is required to setup the virtual machine - as well as on an OCR4all package download.

1. Download the newest VirtualBox version on www.virtualbox.org/wiki/Downloads for your operating system. In the section "VirtualBox ... platform packages" click on "Windows hosts" or "OS X hosts". After downloading the setup file run all installation steps of the VirtualBox software.

2. Download OCR4all package from www.kallimachos.de/ocr4all/vm-download.php. The file name is "OCR4all_0.3.zip". Once the download is complete, unpack the file. The folder at hand, "OCR4all_0.3", will contain all other relevant files.

3. Finally, store the "OCR4all_0.3" folder anywhere inside your operating system, for example in C:\.

Inside "OCR4all_0.3" folder, you will find the following structure:

- "guides" folder, containing detailed instructions regarding software use.

- "image" folder, where the files pertaining to the OCR4all software setup are stored.

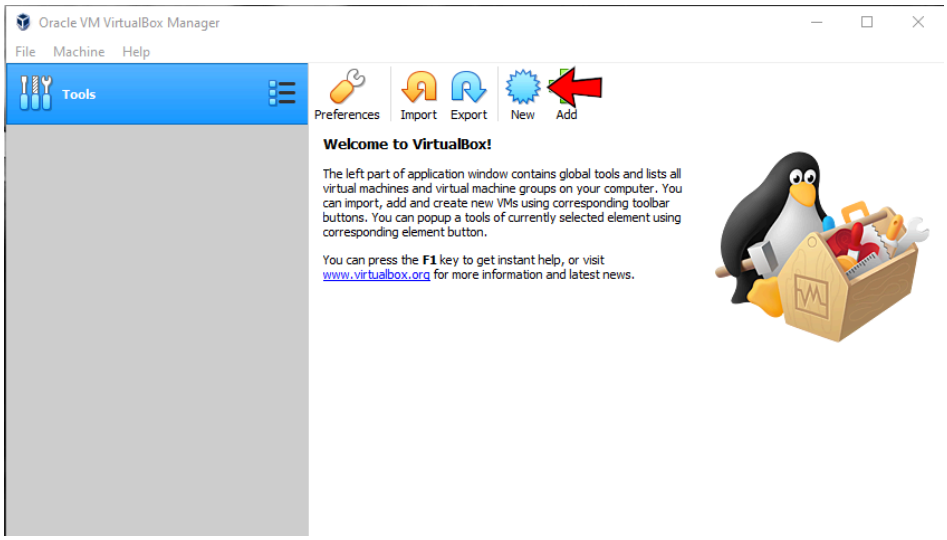
- "ocr4all" folder, containing two further folders, "data" and "models":
  - "data" contains initial testing data needed to operate OCR4all, i.e. scans of several, different prints. This data set is laid out in the exact same way you should later structure your own data and projects: ocr4all/data/*book title*/input/*scans*.
  - On the other hand, "models" contains OCR4all models used during text recognition. Here as well, you will find all the data needed to start using OCR4all at once. All OCR models you will subsequently create while working with the software will be saved inside this folder automatically.

Virtual Machine Setup

In order to setup a new virtual machine for OCR4all, please start the VirtualBox software you downloaded (as shown above).

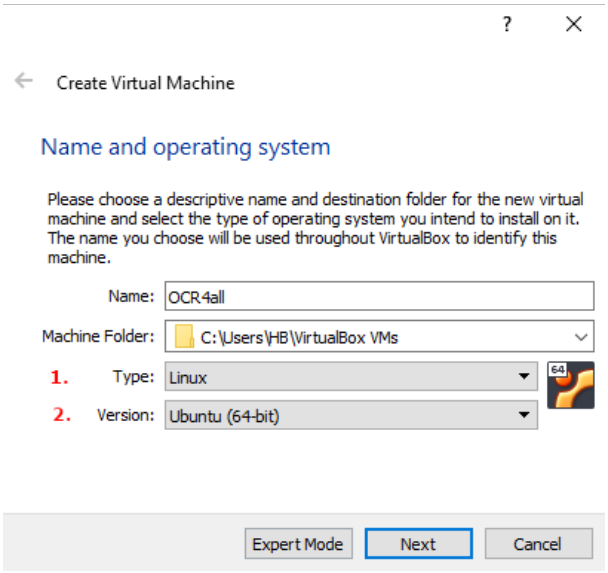
1. Setting up a new machine

Start by clicking on “New”.



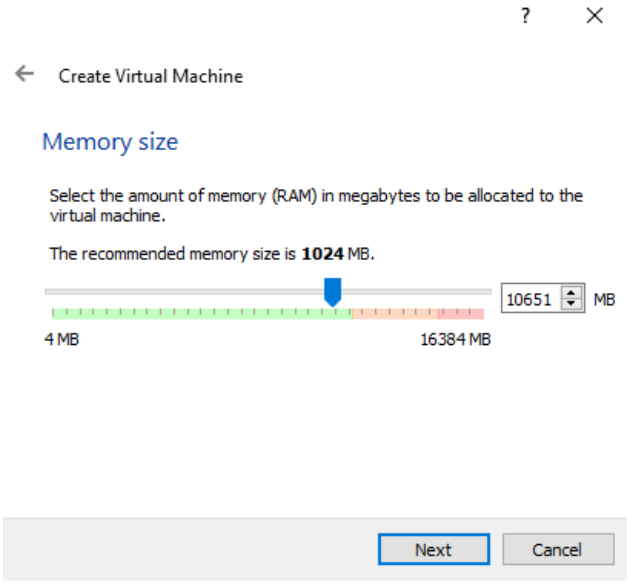
2. Setting up machine parameters

Please note that, regardless of your own operating System, you must select **Linux** in “Type” (1.) and **Ubuntu 64-Bit** (2.) in “Version”. Under “Name”, please type in **OCR4all**. Finally, make no changes to “Machine Folder”.



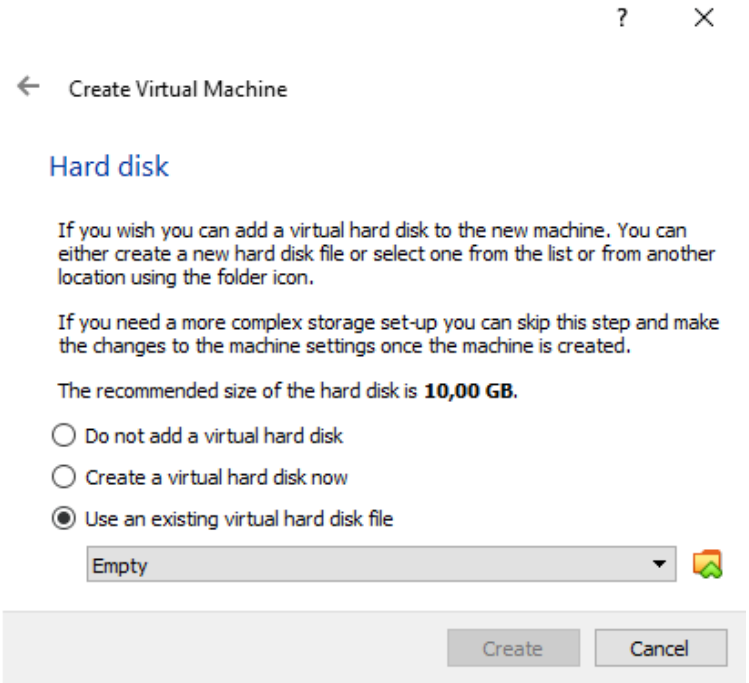
3. Resource management

This step enables you to determine how much RAM (your computer's working memory) you wish to leave at the virtual machine's disposition. We advise you to allocate at most the total amount of your device's main memory *minus* 2 GB. In the following example, this would amount to 14.384 MB.

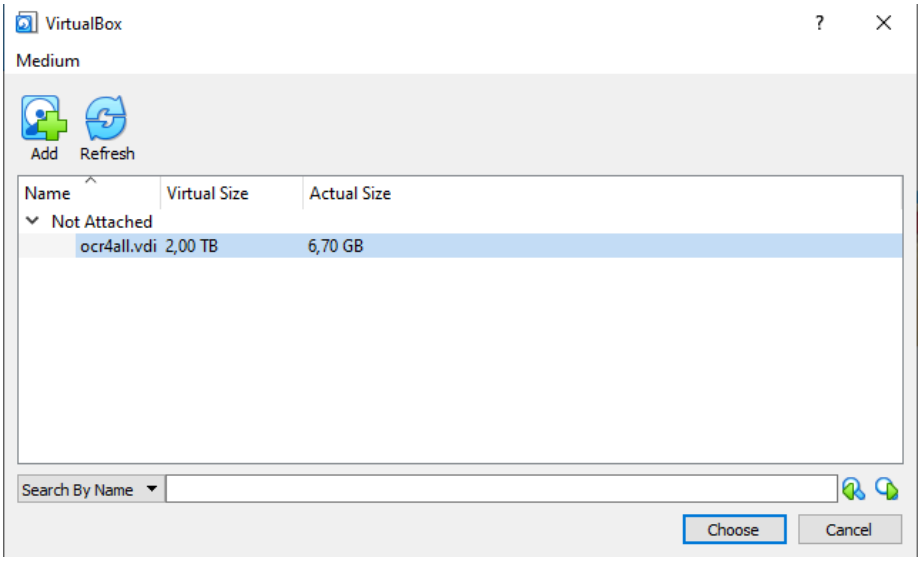


4. Hard drive embedding

Proceed by clicking, as shown below, on “Use an existing virtual hard disk file”. At this point, please note that you will not yet be enable you to create the virtual machine.



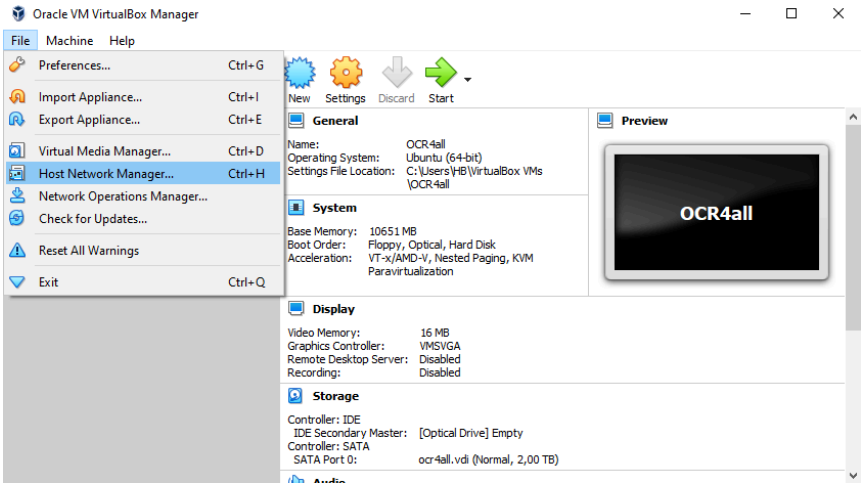
First, you must click on the small folder icon (behind “Empty”) and select the OCR4all image (in “OCR4all_0.3/image”) you previously downloaded. Click on “Add” in order to assign it to the “existing virtual hard disk”.

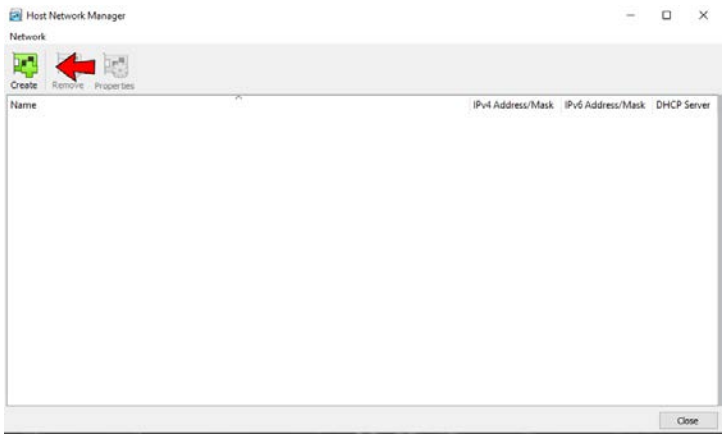


Finally, click on “Choose” to finish the creation of a virtual machine, named OCR4all.

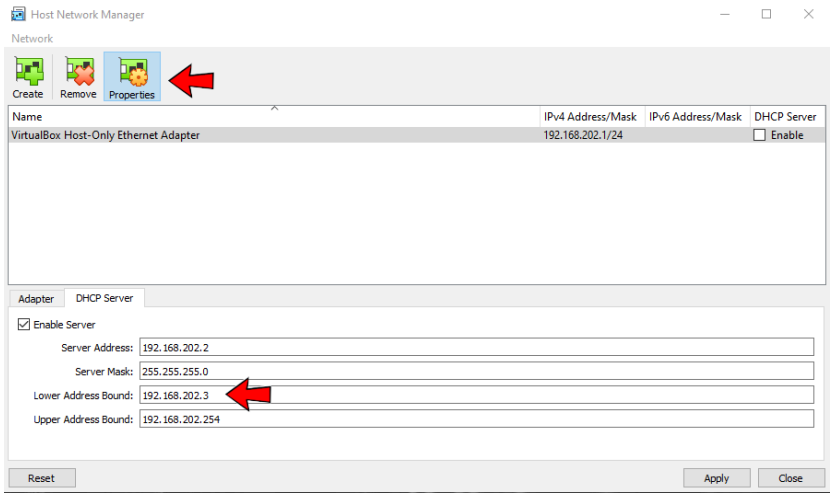
5. Adjust virtual machine settings

In the interest of guaranteeing that the virtual machine operates as smoothly as possible, you will need to make a few more adjustments. The virtual machine and your device must be able to communicate, for which you will need a so-called network adapter. Simply go to menu item **File > Host Network Manager** and click on “Create”.



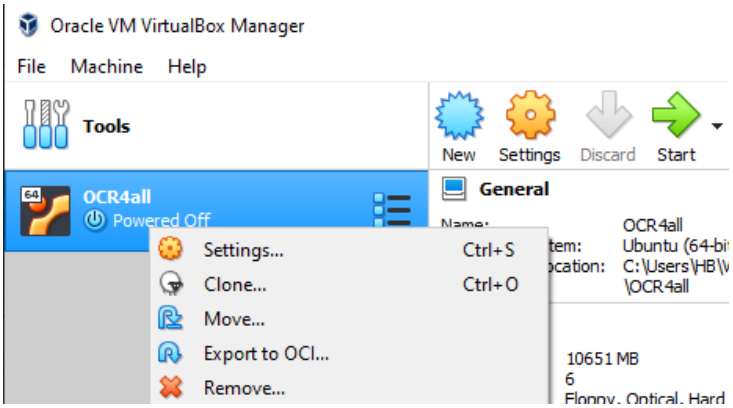


Once the network adapter is setup, click on “Properties”, as shown below. Please make a note of your virtual machine's IP address, which you will need when starting OCR4all. You will find it under “Lower Address Bound”. In the example below, the IP address would be: 192.168.202.3

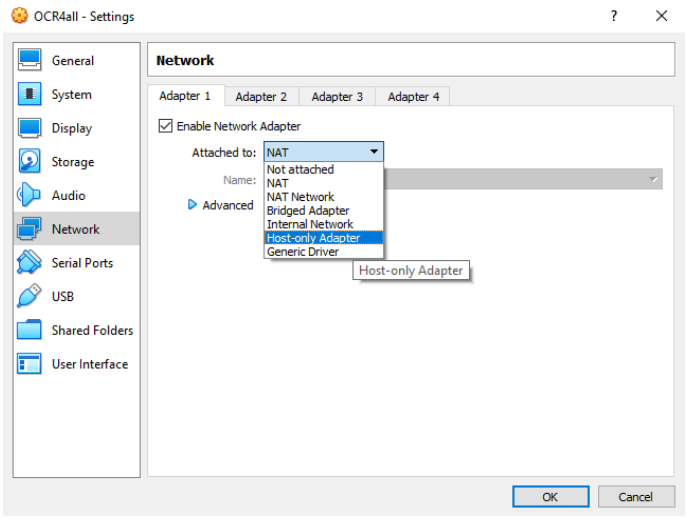


Afterwards, close the Host Network Manager.

To finalise your settings, you will need to go to menu item **Network** inside your virtual machine's settings. In order to do so, click right on the machine and select “Settings”.

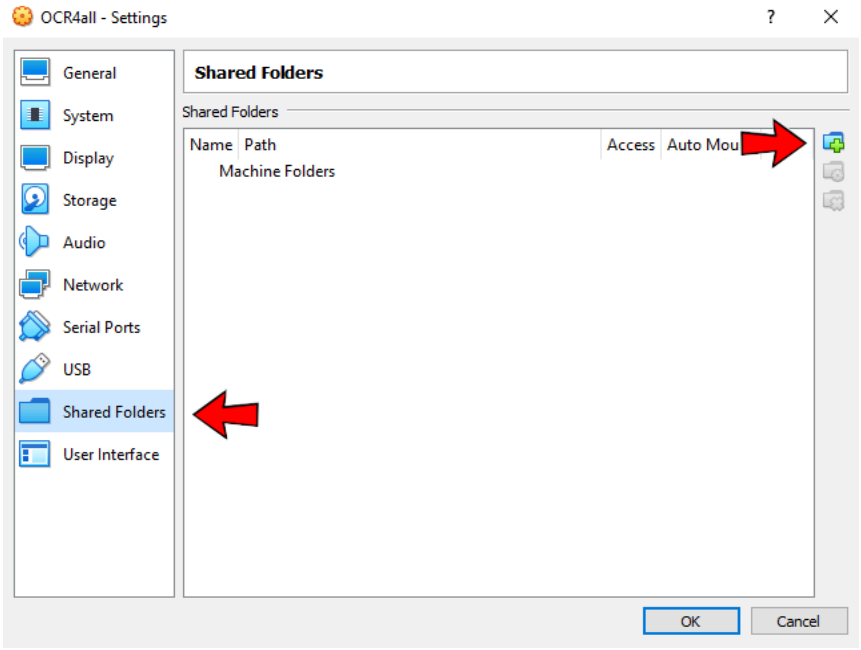


Menu item Network will appear on the left sidebar. Once you've clicked on Network, select "Host-only Adapter", as shown below.



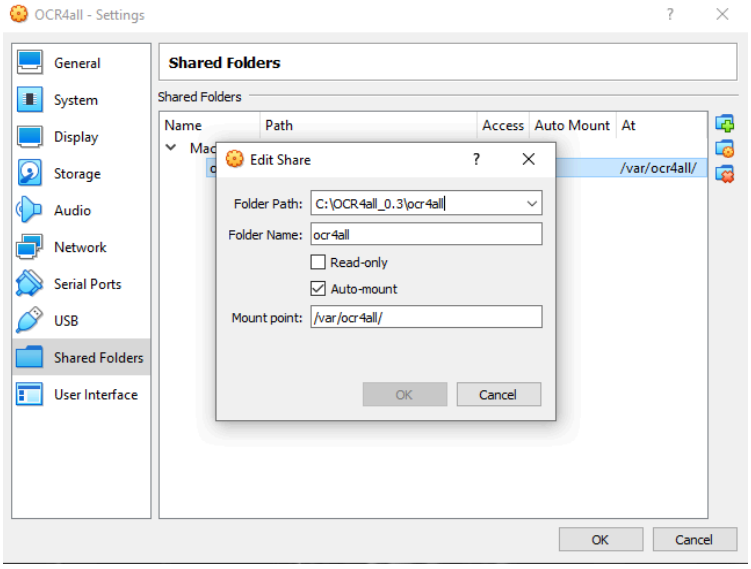
6. Shared folder embedding

In order for a data exchange to take place between OCR4all and your device, you must now create a so-called shared folder. You will find the according menu item in the **Settings** left sidebar as previously described. Once you've clicked on **Shard Folders**, create a new folder by clicking on the small folder icon on the top-right hand side (see below).



Popup menu:

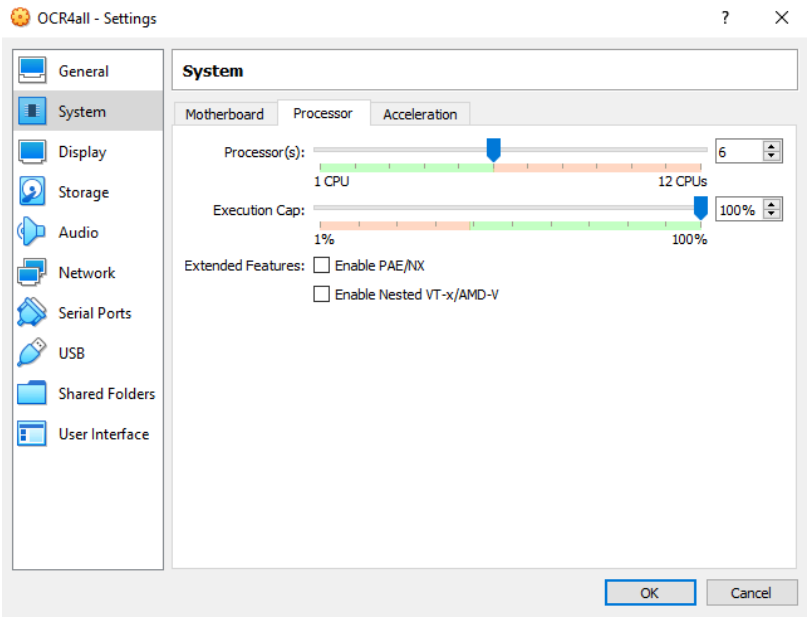
- select "Auto-mount" option.
- under "Mount point", please type in /var/ocr4all/
- the field "Folder Path" must contain the path existing inside your operating system and leading to folder "OCR4all_0.3/ocr4all" you previously created (see below).



Validate with “OK”.

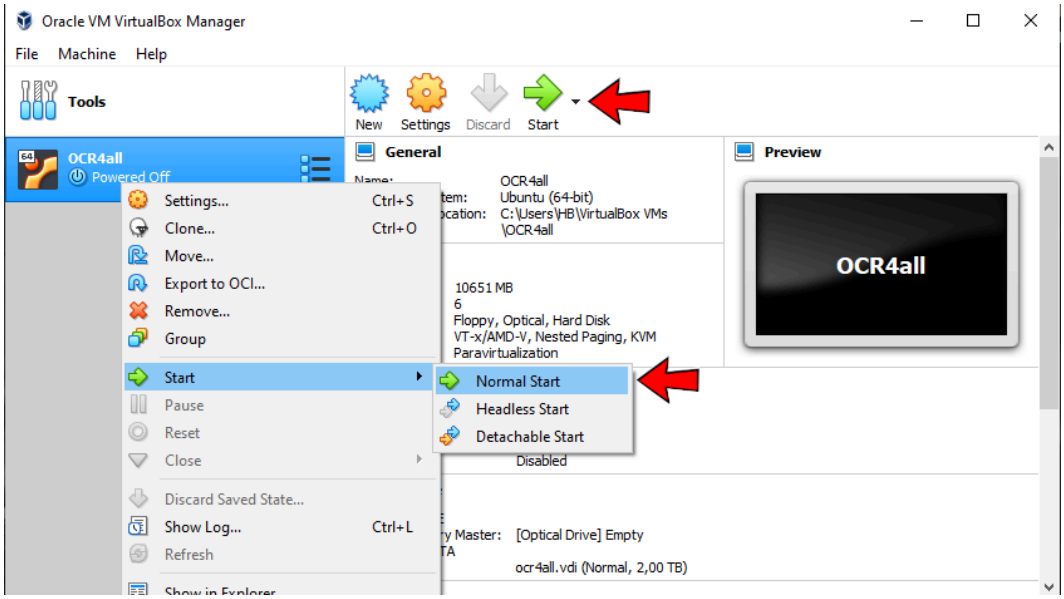
7. CPU core allocation

In the interest of increasing the performance of your virtual machine and of OCR4all, you should finish by embedding further CPU cores: simply go to menu item **System** (inside the OCR4all machine settings, as shown before) and click on **Processor** tab. We recommend you allocate *all but one* of the CPU cores at your disposal.



8. Starting your virtual machine and opening OCR4all

In order to start your virtual machine and to begin working with OCR4all, left click on the virtual machine and click on the green arrow following the steps outlined below.



After you've initiated the start, a new window named “OCR4all” will appear. Wait until the machine is done booting up, which is the case once the inscription “ocr4all login ... Up ... seconds” is visible and steady inside the black field.

You can now start working on your projects with OCR4all by opening your browser (we recommend Chrome) and typing in the following URL:

http://IP address of your virtual machine:8080

This virtual machine IP address is the one you wrote down during step 5. According to our example (p. 6), the URL would be:

http://192.168.202.3:8080

Validate by pressing “Enter”. Your browser should now open the so-called “Project Overview”, which is OCR4all’s homepage.

Contact information, further instructions and documentation on OCR4all

Should you have any questions or problems regarding the software installation, any remarks or suggestions, do not hesitate to contact us, either directly and personally at ocr4all@uni-wuerzburg.de or through a post on GitHub.

You will find additional information about using our software, as well as detailed explanations and step-by-step instructions, e.g. regarding layout segmentation or targeted OCR model training, at OCR4all/getting_started.