WORK IN DENMARK

# Vacancy Scraper Manual

Johannes Mols

jom@workindenmark.dk or johannes.mols@gmail.com

November 1, 2018

# Contents

# 1 Introduction

*Vacancy Scraper* is an application that can organise job postings and the related companies. It is possible to save posted jobs together with their URL to the posting, and then export them in an organised format considering the consultants to the company and the time of the posting. Besides that, the application is able to scrape the company's websites and extract the job titles that are currently online and check if they have already been posted or not.

# 2 Installation

The application can only run on Windows and does not need to be installed like many other applications. It is portable and can, therefore, be placed in any directory. However, it is recommended to place the executable in a dedicated folder in the home directory of the user (together with Documents, Downloads, Videos, ...). It is important that *all* files (e.g. `.dll`, `.xml`, `.pdb`, `.json` files) are placed in the installation folder.

## 2.1 Setting up paths

When starting the application for the first time (double click `Vacancy Scraper.exe`), open the *Settings* tab. You have to define paths of where you want to save the files that the program will be working with.

- **Web Drivers**: this directory contains `.exe` files that are used for scraping websites. The drivers are ideally already bundled together with the application files and can be put in a folder in the same directory. The recommended web driver is Chrome. More drivers can be downloaded using the link below *Browse...*. Below the text box containing the file path, the recognised web drivers are listed. You can select the driver that the application should use with the drop-down list in the *Scraper* box.
- **Resources**: this directory contains four files used to store your companies, completed-, blacklisted-, and open job postings. Select a path and click the blue text below the *Browse...* button saying *Create files*. This will create empty files at the specified location for the program to use. If you have pre-existing resource files, just paste in the path and the files will be recognised.
- **Logs**: this directory contains log files that are created when scraping websites from companies. They give details on what jobs were found, which were considered duplicates and why, as well as outputting error information in case something did not work as intended, which can happen easily when a website only slightly changes.
- **Export**: this path is the default path that is displayed when exporting job postings. It can, of course, be changed but might save some time if the export directory will always be

the same.

- **Browser**: this option let's you change the browser in which clicked links are opened. This can be useful when reinstalling Chrome to make the Scraper work and breaking the standard browser setting in the process.

### 2.2 Scraper Settings

The settings below the paths in the *Settings* tab control certain aspects of the website scraper. You can select the preferred web driver from the drop-down list, provided that there are any available. Job postings will be ignored if they contain a *banned keyword*, which you can define here. Keywords need to be separated by commas. Additionally, you can select if the application should check `jobnet.dk` for possible duplicates, because it can be that someone else posted the same job already.

## 3 Functionality of different tabs

### 3.1 Scrape

**Disclaimer**: Each company needs to have their own scraper individually programmed. Therefore, not many or no companies will be available for scraping. Please contact the developer for a request to add companies to the list. If a company is available for scraping, the check box in the company overview will be checked.

This window displays companies that are enabled to be scraped in a list. They can be individually enabled or disabled by clicking the check box in the list. You can start the scraping by pressing the *Run* button on the top. The application will open and close several browser windows whilst scraping the websites. The text box on the lower half of the window will display information about the progress. Detailed information can be found in the log files. Once the process is done, the found vacancies are added to list of open vacancies and can be found in the *Vacancies* tab.

It is likely that you will get an error message when trying to start the Scraper. This is because Chrome and Firefox are installed through the Software Center and the program can not handle such an installation properly. If you want to make it work, you will have to uninstall Chrome or Firefox (Chrome is recommended) in the Software Center and download the normal Chrome installer from Google's website. You don't need to have administrator rights to do this, simply click *No* when prompted for an administrator login, and then select to install anyway. **You will not have a standard browser after doing this**, meaning that there will be errors when opening links in applications. Please ask an administrator to reset your standard browser. To circumvent this in the application, there is an option in the Settings tab to select any of the installed web browsers. Links that are opened within the application will now use that browser.

## 3.2 Vacancies, Blacklist and Done

These three tabs all work the same way but serve different purposes.

The *Vacancies* tab displays jobs that have not been posted yet but will be in the future. Normally, only the Scraper will add jobs here, but you can still use it manually by clicking the *Add* button. Deleting vacancies can be done by selecting one or more vacancies in the list (control-clicking, shift-clicking, Ctrl+A, ...) and then clicking the *Delete* button. Vacancies can be blacklisted as well as marked as completed using the same way. The vacancy will then be removed from the current list and added to the new one, updating the timestamp to the current time.

The *Blacklist* tab works exactly the same way, with the exception that you can restore blacklisted vacancies to the open-vacancies list by clicking the *Restore* button. You can not directly mark blacklisted vacancies as completed.

The *Done* tab works exactly the same as the *Blacklist*. All completed job postings will be saved here.

### 3.2.1 Manually adding vacancies

You can manually add vacancies to any of the lists by clicking the *Add* button on the top.

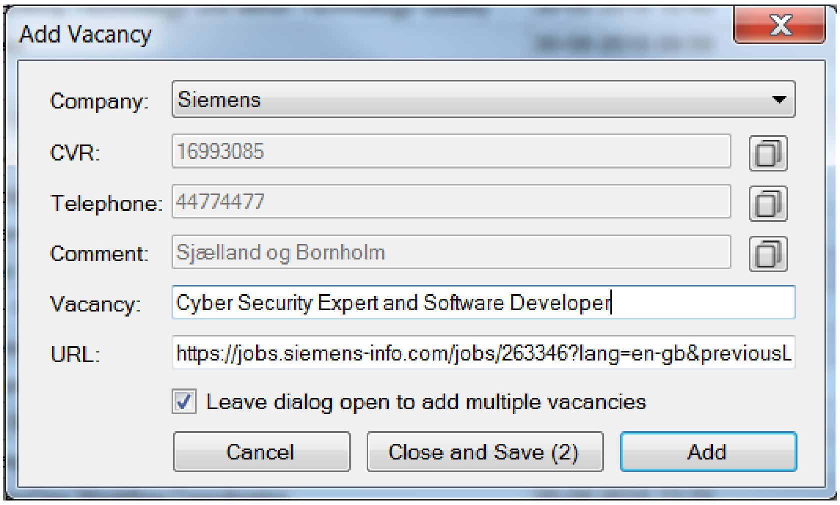


Figure 1: Adding a vacancy

A new window will open where you can select the company of the position from the drop-down list. This list is an alphabetically sorted list of all the companies that you entered in the

*Companies* tab. The drop-down list can be operated by pressing a letter on the keyboard as the first thing after opening the window. This will select the first company starting with that letter. You can go through the list with the arrow keys or by simply opening the drop-down list with the mouse and selecting the company.

There are three greyed-out text boxes below the company list, which serve the purpose of showing important information you will most likely need when posting a job. This includes the CVR and telephone number of the company, as well as any comments you may have added to the company. You can copy the contents of those text boxes by clicking the small button to the right of each text box.

The text box titled *Vacancy* needs to contain the title of the job posting, and the *URL* is the link to the job posting on the company's website.

You can select to add multiple vacancies at once by clicking the check-box titled *Leave dialog open to add multiple vacancies*. You can then click the *Add* button to add the vacancy to the list of vacancies you want to add at the same time. A number indicating the number of vacancies that will be saved will appear on the button in the middle titled *Close and Save*. You can use this button to close the window after adding multiple vacancies and save all of them. Pressing the *Cancel* button will close the window and not save any of the added vacancies.

If you are trying to add a vacancy that has already been added, a warning will appear and ask if you still want to add it. Only the company and vacancy title will be considered for this, not the URL.

### 3.2.2 Editing cell values

Editing vacancies can be done by double clicking a cell and typing a new value. Pressing Enter or clicking on a different cell will save the change. Not all cells are enabled for editing (e.g. the timestamp).

### 3.2.3 Copying cell values

You can copy the values of an entire row by pressing `Ctrl+C` while having one or more rows selected. Single cell values can be copied by going into the edit mode (double click on cell), selecting the relevant text, and then right-clicking on the text, and then clicking *Copy*.

### 3.2.4 Following URL's

The URL's in the list can be opened in the browser by control-clicking the green underlined URL.

#### 3.2.5 Searching

You can search for vacancies by typing something in the search box in the top right corner of the tab. The list will update and only show vacancies that contain the search term in their title.

If you want to only display vacancies from a single company, you can start the search term with the **@** symbol and then type the company name. You don't have to type the full name, the list will update live and show you vacancies from companies that contain the search term. This way, you can also display vacancies from multiple companies with similar names.

Additionally to filtering results by company, you can search for specific jobs for that company. For this, you need to start a new search term by adding the **&** symbol after the company search term and then type the title of the vacancies you want to search for. It is important to not leave a space between the company search term and the vacancy search term. An example might look like this: `@Novo&Scientist`

Figure 2 shows the example displaying vacancies that are from a company containing the term "Novo" and containing the word "Scientist" in it's title.

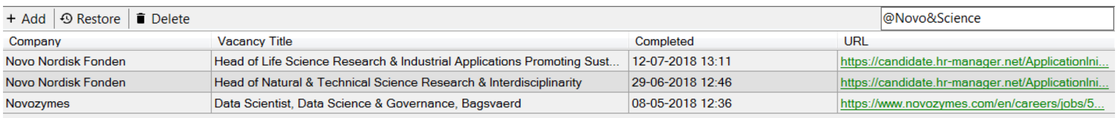


Figure 2: Search example

#### 3.2.6 Sorting

You can sort any of the lists by clicking on the column header. The sorting order will switch between ascending and descending with every click.

### 3.3 Companies

The *Companies* tab works similarly to the other table-based tabs. It shows all available companies and relevant information to them. You can add and delete companies by using the buttons on top.

Companies are grouped by consultants to the companies, which is important to grouping vacancies when exporting them (companies with the same consultant(s) will be exported together). The table shows groups in alphabetical order and orders the companies within a group by oldest to newest (new companies will always appear at the bottom). If this order needs to be changed, it has to be manually done in the resource file.

Companies can be assigned a CVR, P-Number, telephone number and comment. This information will appear when adding vacancies manually and helps entering relevant information.

The URL can be opened in the default browser by control-clicking the green underlined text. If the URL is not green and underlined, it is not a valid URL (check if it starts with `http://` or `https://`).

The column called *Scraping* is an indicator whether there exists a scraper or not. It can't be manually be checked or unchecked because each company requires their own custom scraper that needs to be programmed separately.

Searching for companies is done by typing a search term in the search box in the top right corner. The list will update with companies containing the search term in it's name.

You can sort the list by clicking on the column headers. The list will switch between ascending and descending with every click.

## 3.4 Export

The *Export* tab serves three different purposes: exporting, importing and synchronising vacancies.

### 3.4.1 Export

Vacancies can be exported to a `.txt` format, listing each vacancy grouped by company.

The data source can be selected with the drop-down list. You can export open, blacklisted and completed vacancies with this method.

The date range determines which vacancies will be exported. Only vacancies that have been added/blacklisted/completed during within this range will be included. The current week is selected by default.

In the next step, the consultants can be selected or deselected. If you only want to export vacancies from companies that have a specific consultant, you can choose that here.

The directory is the path to the folder where the exported file will be saved. You can paste a path into the text box or click the *Browse...* button and select a directory.

The file name is the name of the file that will be exported.

After clicking the *Export* button, you will be asked if you want to open the file location of the exported file. If choosing yes, the file browser will open and show the folder.

### 3.4.2 Import

You can import vacancies and companies to your own data. This can be useful if you're setting up the application and have backed-up data or receive them from a colleague.

The application works with `.json` files, meaning that the files you want to import have to be of that format. The application can only read files that also have been created by the application.

You can choose to import files for any of the four resources. It will check for duplicates with your local files and ask if you want to add duplicates, if there are any.

Next to the four buttons, you can see how many vacancies/companies you have in a category.

### 3.4.3 Synchronise

You can synchronise your local files with a Google Drive account to back up data and prevent potential losses. It can also be used by several people to always work on the same version without moving files from computer to computer.

You can either upload or download files to the Drive, overwriting everything, locally or in the Drive. You can choose to only synchronise specific files, or all.

Clicking the *Upload* or *Download* button for the first time will open the Google website in the browser, asking you to log in with a Google account and give permission to the application to manage your Drive. This has to be done only once.

If you would like to remove the Google account from the application and potentially choose a new account, you have to locate the installation folder of the application and delete the `token.json` folder.