extract Text from a WEB Page

this small program extracts data from WEB pages. Condition, the source must not be password protected.
The text parts found in the page are entered as individual lines in a list.
If a vMix system is connected, the single lines are transferred to a gtzip title.
The program and the gtzip Title can be downloaded at the end of this article.
The program is a NET Framework 4.8 application and needs to be installed.

To get this program running you will need some computer knowledge.

BASIC REQUIREMENT: Chrome must be installed and chromedriver.exe must be present in a known directory of the computer, you are running this program. The file „chromedriver.exe“, is used to read the data from the web pages.
If chromedriver.exe does not match the installed version of chrome or is not present or the location of chromedriver.exe is not entered correctly in to this software, the program will not work.
Chromedriver downloads
Chromedriver betas, downloads for actual Chrome Browser! (mine runs 118.x)

If you don't have any experience how to install chromedriver.exe in the desired version to a location you know and then enter this path into this program, then you need to take care of it first. Google or an experienced computer expert in YOUR environment can help

IMPORTANT: Since JavaScript’s take time and especially decoding web pages, the minimum interval of autoupdate is 5 seconds. Autoupdate works only on the 30 individual ID's. Please note that too fast and regular query could be blocked at some point on server side.
Be aware that extracting values from a web page is not comparable in time to a data stream from a time display. It all takes a little time and the willingness of you to exercise a little patience.
In the top line of the program window you have the possibility to choose between two search modes, start with Fast Search.
Deep Search gives the test Chrome browser, running in the background, the possibility to build the page, but this may take a little more time depending on the code of the page, but also because of the speed of your Internet line. You can set this time in the field next to Deep Search. Sometimes a waiting time of one second is enough to find the searched values with Deep Search. Please note that too fast and regular query could be blocked at some point on server side.
You can also click on search process visible in the upper right corner. Then you can see if the text browser has built the page. But this only makes sense if you could interpret the results with the Inspector.

- 1. First, enter the address of the website from which you want to fetch the data.
- 2. Then click on the button „click here to try to find from the web page all the IDs and their corresponding values“.
- 3. Drag the desired values from the left list into the 30 text boxes on the right. (maximum 30)
- 4. Click on „click here to extract only selected ID's“, these values will then be selectively read and transferred to vMix, if desired and possible.

The program also reliably searches for ID's which are only generated at runtime of the web page (Deep Search). That means it can also find ID's which are defined by stylesheets and generated at runtime by Java Script.
But if this program can't extract values, which you can see on the web page, it can have different

reasons. Most of them are that the output of the data you want is running in an iFrame, are retrieved from an external server at runtime and mapped as classes or data.
In order to explicitly read such pages, a specific analysis of the page's source code and programming tailored to this result is needed. Or you can use paid online services like browse.ai, ScrapingBee or fivetran. (this list is incomplete)

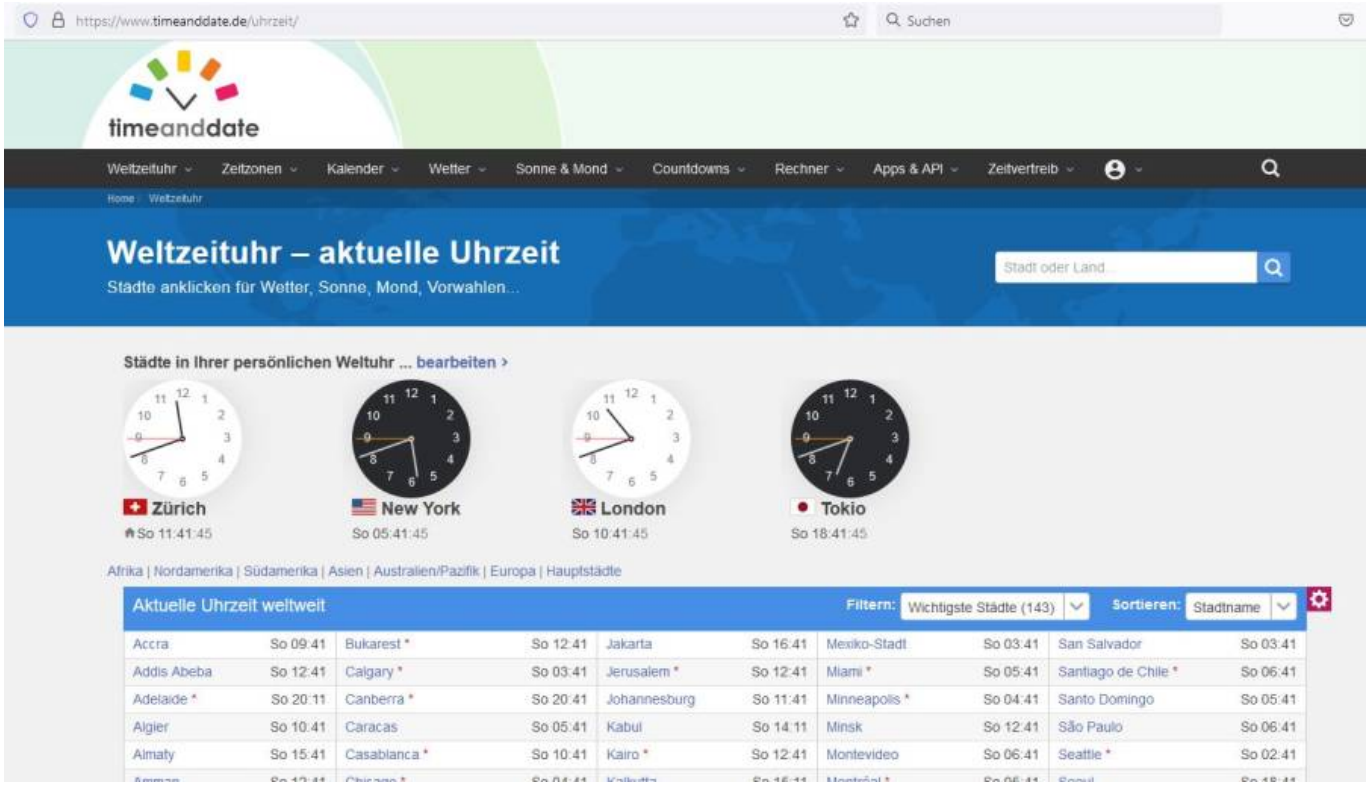
If the vMix transfer function is enabled, but vMix is NOT running, the program will respond very slowly. So please turn off the vMix function when vMix is not running.

In the vMix section of this program, the name of the vMix GT-Title can be changed.
If you want, you can send any ID's directly to a GT-Title of yours. ID1 corresponds to GT-Title Text1.text, ID2 = Text2.text etc. Please note that the name is case sensitive
If 30 ID's are too few, you can run multiple instances of this program and send the results to different GT-Titles.

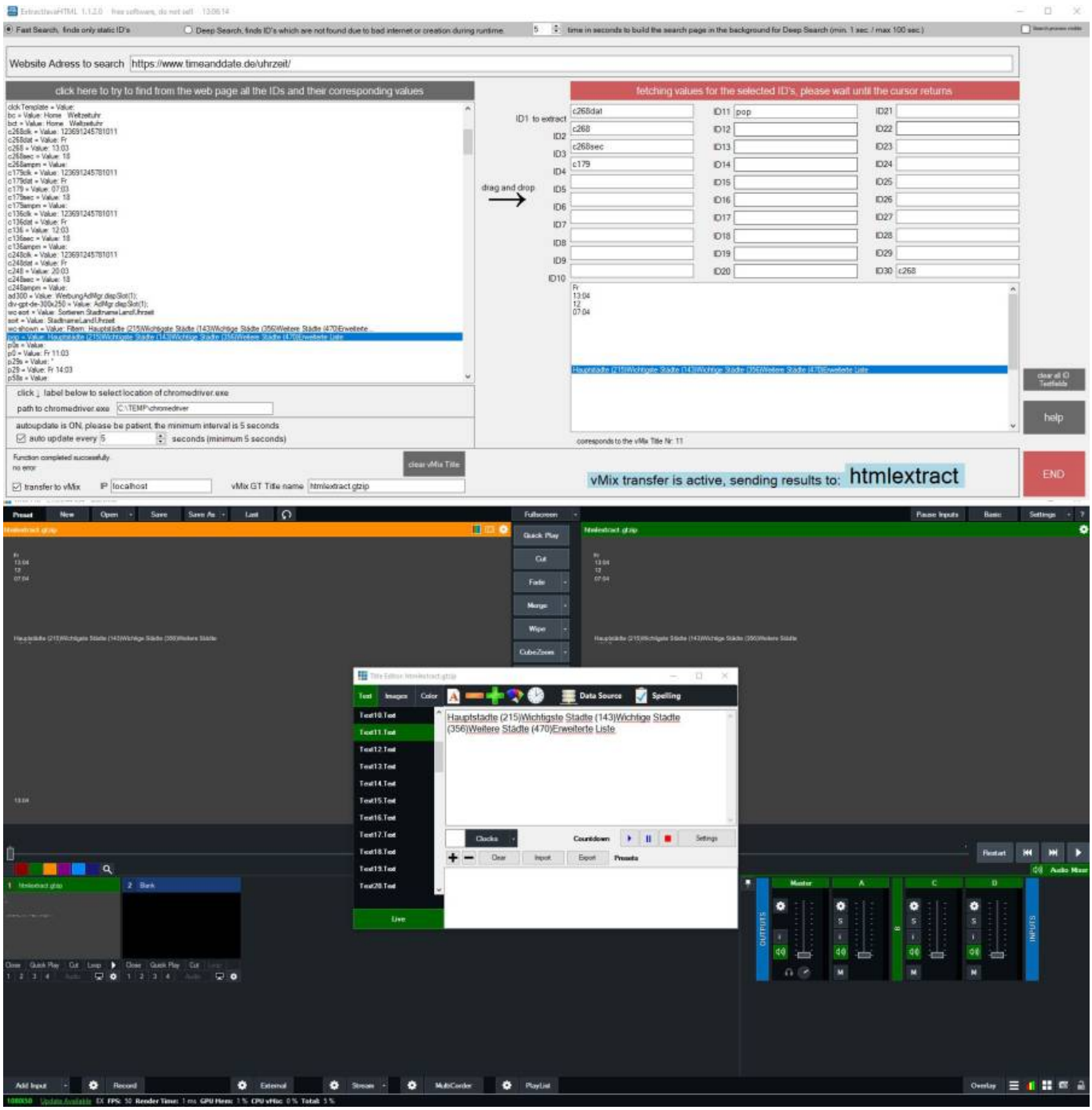
All fields with variables, with the exception of autosave and transfer to vMix, are automatically saved when a value is changed or when the program is terminated and restored at startup.
Autosave andOr transfer to vMix must be explicitly switched on by hand at every startup.

I tried to pack as much error handling as I could into the program. But if there are any errors in the program, I apologize for them and explicitly exclude any liability.

from here:



to here:



Download zipped Windows setup program and the vMix title here:
ExtractHTML

From:
https://tvcrew.ch/vmix/ - **vMix Wiki Deutsch**

Permanent link:
https://tvcrew.ch/vmix/doku.php?id=extract_text_from_a_html_web_page

Last update: **2023/10/20 20:27**