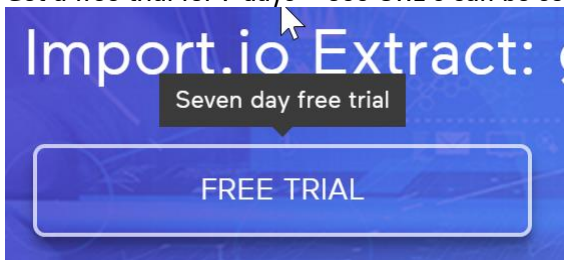


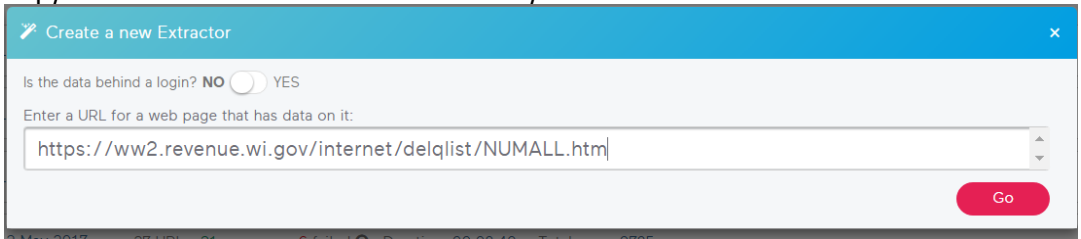
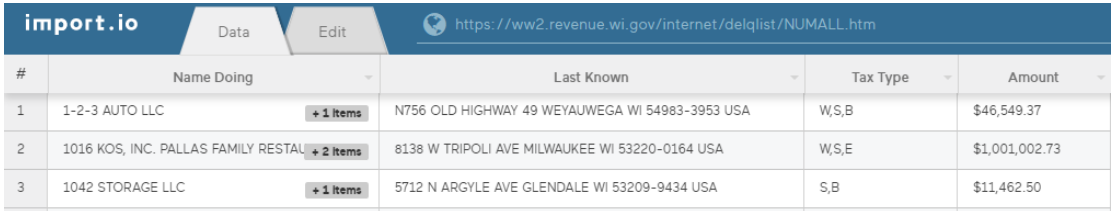


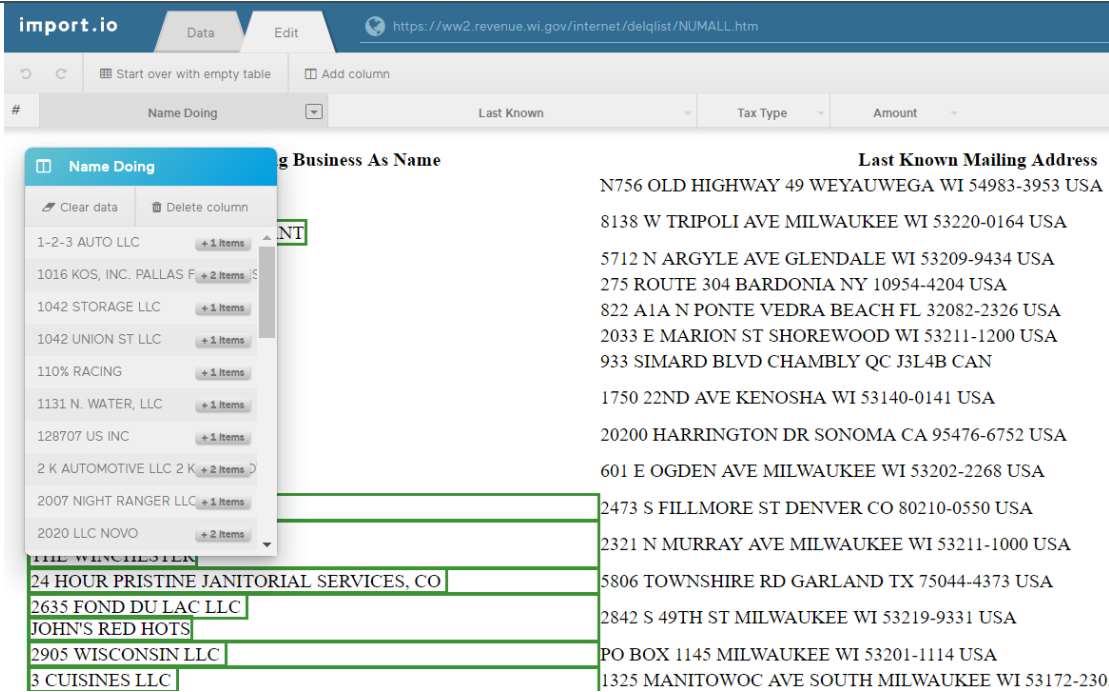
MEMO

| What | Why | How |
|--|--|--|
| Import.io - scraping without programming | <p>Import.io is a highly automated web-scraper, building on AI, finding patterns in websites.</p> <p>Many scraper-tasks can be solved by copy URL into the program.</p> <p>Registrate and get a seven day free trial.</p> <p>There has a year ago been a special agreement with journalists, but now it is demanded that you make an individual negotiation. And they are bot easy to work with.</p> <p>Despite it is very automated, there is still al lot of demanding issues you need to know or learn.</p> | <p>Main-site: https://import.io/</p> <p>User-guide and tutorials: https://help.import.io/hc/en-us</p> <p>Evt. discounts for journalists: kortlink.dk/nxv9</p> <p>Community: http://community.import.io/</p> |
| Free trial | <p>Get a free trial for 7 days – 500 URL's can be scraped.</p>  <p>The image shows a blue banner with the text 'Import.io Extract: Seven day free trial' and a large white button with the text 'FREE TRIAL'.</p> | |
| Go to Dashboard and click on "New Extractor" | <p>Clicking on New Extractor makes it possible for you to build your scraper.</p>  <p>The image shows a red button with a white hand icon pointing to it and the text '+ NEW EXTRACTOR'.</p> | |

MEMO

| What | Why |
|-------------------------------------|---|
| Inspect the web-site to be scraped. | Before building a new scraper inspect the data on the website, you want to scrape. Which data do you need? Is there a structure in the way the URL's are build. Do you need go down on more levels – then you need to extract the url's first – and then use them as input in another scraper? |
| Start New Extractor | By clicking on New Extractor, you start a new job:  |
| Paste the URL into the system | Copy the basic url for the data-site into the system and click Go:  |
| Edit scraper | Sometimes it start with the option of a blank scraper – other times it suggest the columns and rows, and you get a suggestion, which you only need to edit a bit – or not at all. In this case the suggestion is this – and this is fine:  If you are not satisfied, you can edit the scraper. You can rename column, add or remove column, and you have more advanced options. |

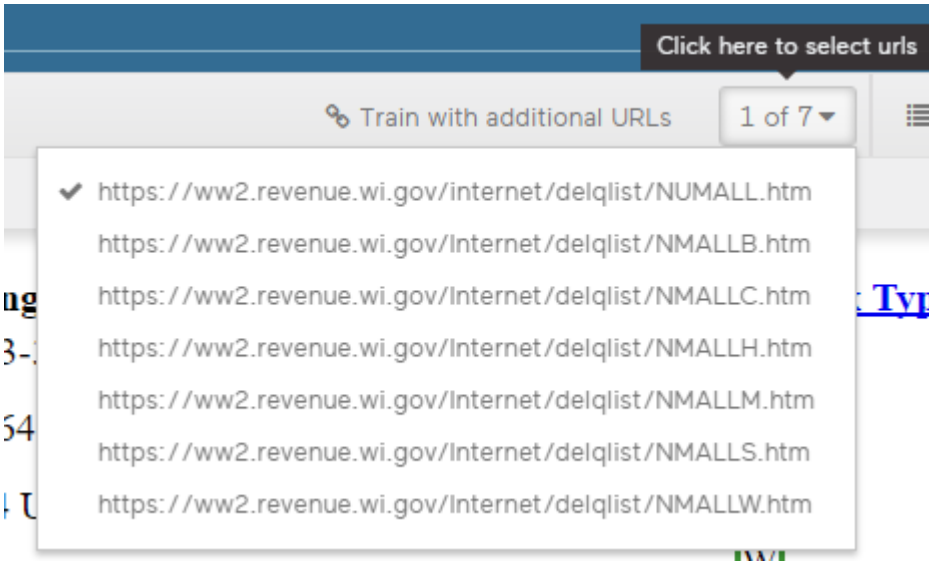
MEMO



The screenshot shows the import.io interface with a table of business data. A dropdown menu is open for the 'Name Doing' column, showing a list of businesses. The table has columns: #, Name Doing, Last Known, Tax Type, and Amount. The data is as follows:

| # | Name Doing | Last Known | Tax Type | Amount |
|---|--|------------|----------|--------|
| | 1-2-3 AUTO LLC | | | |
| | 1016 KOS, INC. PALLAS F | | | |
| | 1042 STORAGE LLC | | | |
| | 1042 UNION ST LLC | | | |
| | 110% RACING | | | |
| | 1131 N. WATER, LLC | | | |
| | 128707 US INC | | | |
| | 2 K AUTOMOTIVE LLC 2 K | | | |
| | 2007 NIGHT RANGER LLC | | | |
| | 2020 LLC NOVO | | | |
| | 24 HOUR PRISTINE JANITORIAL SERVICES, CO | | | |
| | 2635 FOND DU LAC LLC | | | |
| | JOHN'S RED HOTS | | | |
| | 2905 WISCONSIN LLC | | | |
| | 3 CUISINES LLC | | | |

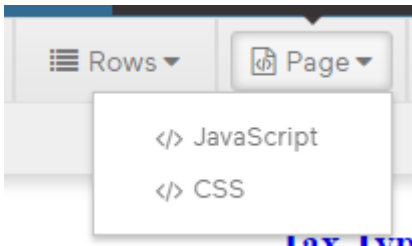
Here the same website and data is shown after clicking on Edit, which gives a possibility to edit the content of the data in the fields. Here no edit is needed.



The screenshot shows the import.io interface with a list of URLs to train with. The list includes:

- https://ww2.revenue.wi.gov/internet/delqlist/NUMALL.htm
- https://ww2.revenue.wi.gov/Internet/delqlist/NMALLB.htm
- https://ww2.revenue.wi.gov/Internet/delqlist/NMALLC.htm
- https://ww2.revenue.wi.gov/Internet/delqlist/NMALLH.htm
- https://ww2.revenue.wi.gov/Internet/delqlist/NMALLM.htm
- https://ww2.revenue.wi.gov/Internet/delqlist/NMALLS.htm
- https://ww2.revenue.wi.gov/Internet/delqlist/NMALLW.htm

Here I have added several URL's to test that they all work in the scraper.



The screenshot shows the import.io interface with a dropdown menu for JavaScript and CSS. The menu is open, showing the following options:

- JavaScript
- CSS

Sometimes it is necessary to turn of JavaScript and CSS to come down to specific possibilities

MEMO

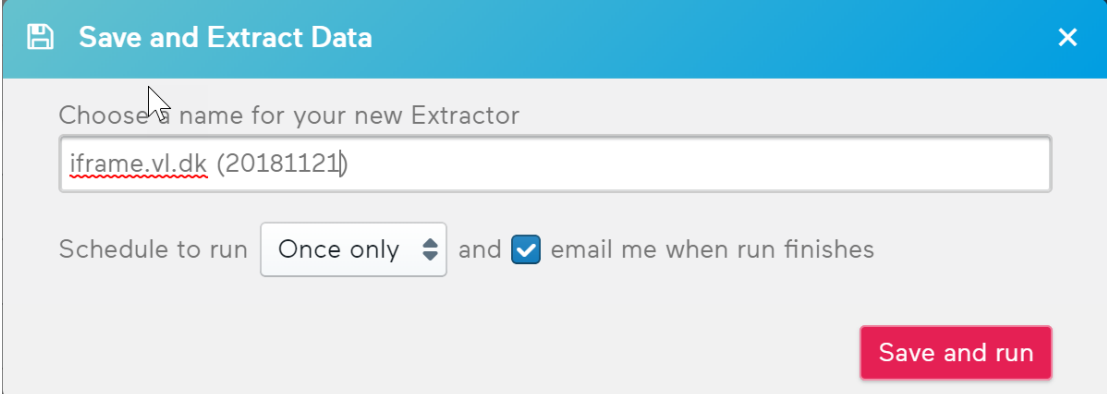
Finish the scraper

for selection.

Sometimes you need even more manual settings to define the right extraction.

Click Extract data from website when the scraper is ready:

 Extract data from website



Save and Extract Data

Choose a name for your new Extractor

iframe.vl.dk (20181121)

Schedule to run and ☒ email me when run finishes






Save and run




If you click in the name of the scraper, in this case iframe.vl.dk you can change it to a more relevant name. Be sure to name scrapers with relevant names. I will call this VLDetail20161115, because it extract the detailed info on Danish members of a business-network called VL (Virksomheds Ledere).


I will also from here be able to run it, or click on edit, to go back in the edit-mode. I can duplicate the scraper and then edit a new version, while keeping the old.





Run and check the scraper

After clicking on Run URLs I get this result:

 Settings
  Run history
  Integrate



 15 Nov 2016 URLs: 1/1 00:00:06s 35 rows
 



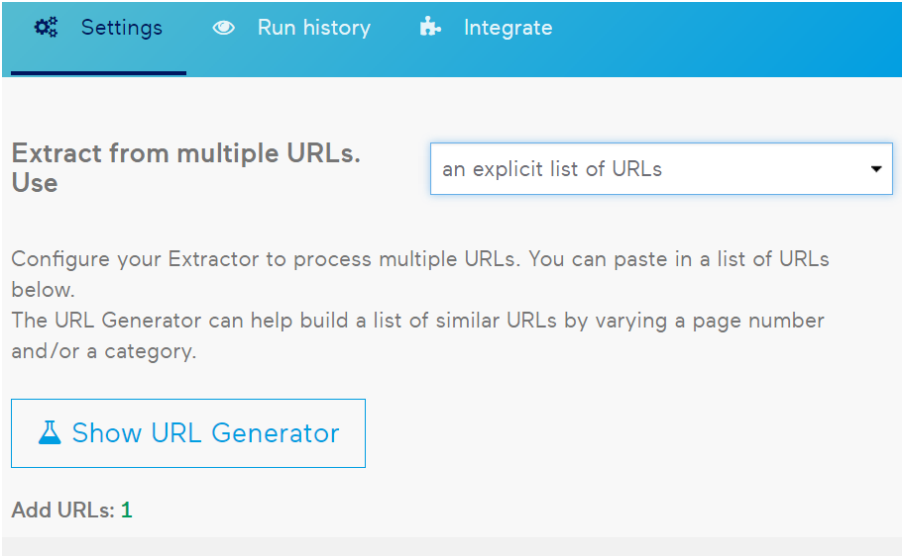
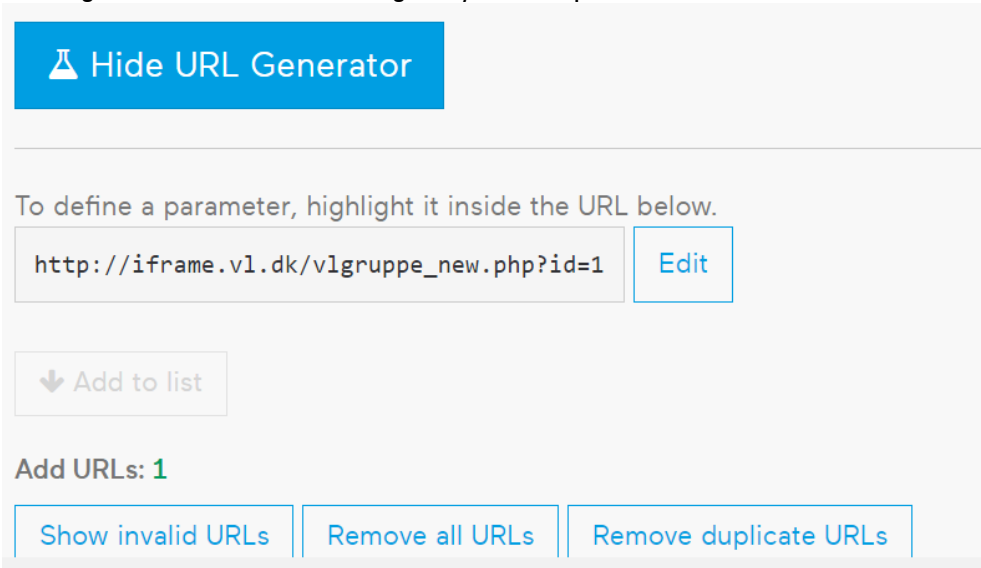
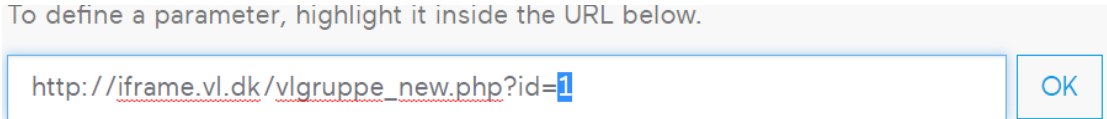
| Title | Firstname | Lastname | Company |
|-------------------|-----------|------------------|----------------------------|
| Bestyrelsesmedlem | Birgit | Aagaard-Svendsen | Axis Offshore m.fl. |
| President & CEO | Jens | Birgersson | ROCKWOOL International A/S |

The run has been performed 15 Nov 2016 and only the provided URL has been extracted with

MEMO

| | |
|----------------------------|---|
| | <p>the result of 35 rows. I can download the data as a CSV-file or JSON. CSV is the best option for opening it in Excel. Be aware it uses US-formats of numbers. Clicking on the eye gives a preview of data. If a link do not load, the system will retry a number of time. If there still are problems, you can download a log-file, by clicking at the text-ikon. In this case, the log-file has no problems. It looks like this:</p> <pre>' Time,"Url","Success","Error Type","Error" 2016-11- 15T17:02:11.982,"http://iframe.vl.dk/vlgruppe_new.php?id=1","1","",""</pre> |
| Problems with URL's | <p>If some URL's just don't get extracted, consider to build a new extractor for them, instead of continuing to try to solve it inside the first scraper. Or download them by Table Capture or another solution. Just get it solved as fast as possible.</p> |

MEMO

| What | How |
|--------------------------|---|
| More than one URL | <p>If data has to be scraped from more than one single webpage, you have three options.</p> <ol style="list-style-type: none"> 1. You can do a scraper to retrieve all URLs, run it first and use its output as input to the actual scraper. 2. You can form the many URLs in a spreadsheet and copy them in 3. You can generate the URL's by clicking on Show URL Generator. The possibility will appear by clicking on "Settings":  |
| Generate URL's | <p>Clicking on Show URL Generator gives you this option:</p>  <p>Click Edit, highlight the relevant parameter, In this case, "1" and click OK:</p>  <p>It may take some time and require you to try sometimes, before you can edit it.</p> |

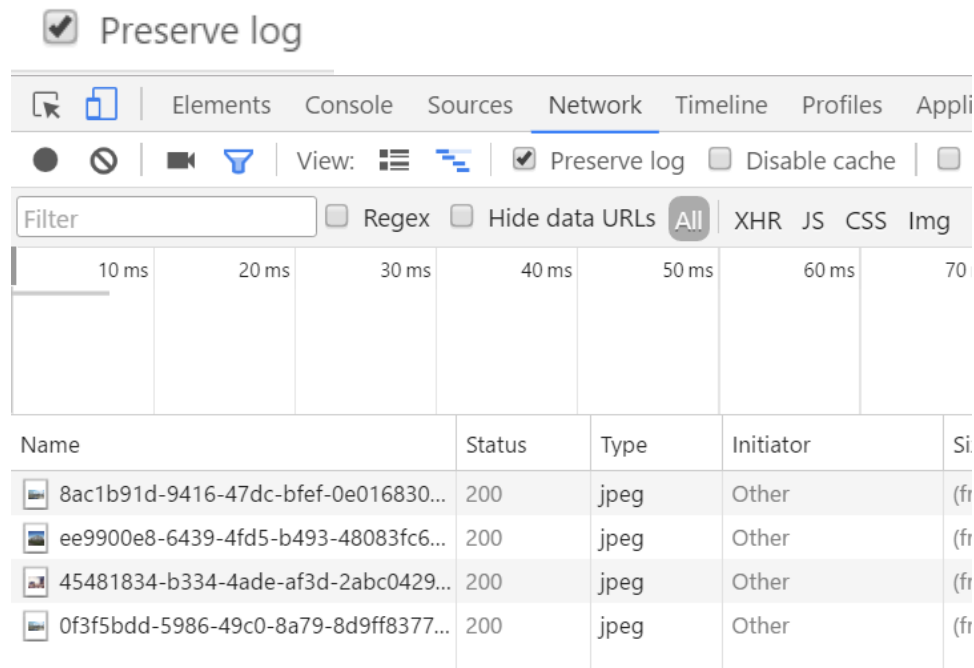
MEMO

| | |
|--|---|
| | <div><div><div>×</div><div>Parameter-1</div></div><div><div>Range of numbers</div><div>1</div><div>to</div><div>121</div><div>step</div><div>1</div></div></div> |
| | <div>Number of generated urls: 121</div> <div><div>http://iframe.vl.dk/vlgruppe_new.php?id=1 http://iframe.vl.dk/vlgruppe_new.php?id=2 http://iframe.vl.dk/vlgruppe_new.php?id=3 http://iframe.vl.dk/vlgruppe_new.php?id=4 http://iframe.vl.dk/vlgruppe_new.php?id=5 http://iframe.vl.dk/vlgruppe_new.php?id=6</div></div> <div>Here is a series of numbers ranging from 1 to 121 with an increment of 1 each time. Add the generated URLs and remove duplicates. Then, "Run URLs".</div> |

MEMO

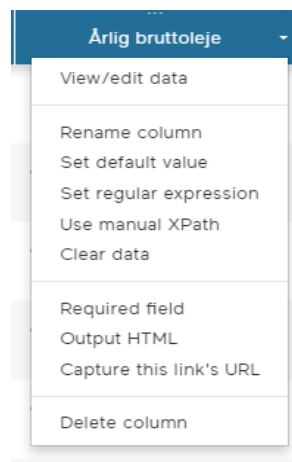
Find hidden URL's

If a page loads as you scroll down, set Inspect, select Network, and click Preserve log, then the page records what happens when you scroll down. From here you can probably identify the underlying links, and then build a scraper that takes these links one by one. This is also the method if you can't find the link for clicking on next for results on more than one page.

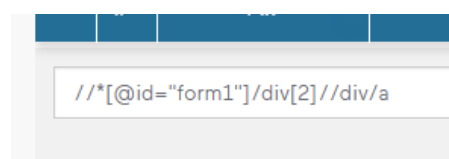


Manual XPath

Clicking on the drop-down menu in Import.io gives you a number of choices. Should you extract a link, it must be on Output HTML. To insert XPath, click Use manual XPath.



In this field, enter the code for XPath.

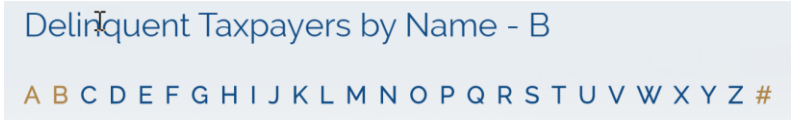
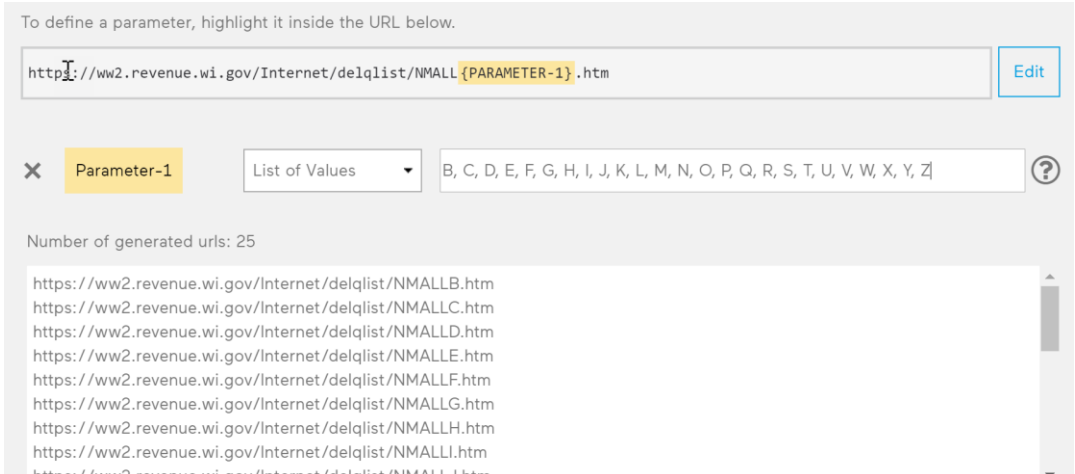


See also the manual for XPath:: <https://guide.import.io/using-manual-xpath.html>

MEMO

| What | How |
|--|--|
| Optimize bug fixes | There may be many bugs in scrapers. Here we review some of the most common - and the method of correcting them. |
| Train Multiple URLs | <p>When you create a new Extractor, then add urls to test that it works on multiple pages. If not, add the missing data - and deselect the wrong one. Possibly you need to start from scratch choosing data.</p> <p>The extractor is usually good at learning what's to include and what's to exclude. You might train it on pages with fewer answers so it don't use too much memory.</p> |
| Error-codes | Review error messages to see why they are failing. Then go to the specific page to resolve the error. |
| Split scrapers | If multiple URLs always fail, make a special scraper for these errors - so you do not try to solve the impossible inside one scraper. |
| Use manual XPath | If it is impossible to train the selection automatically, try the manual XPath. |
| Disconnect Styles and Scripts | Some webpages allow access to data only if you disconnect styles and/or scripts. Try it out. This is especially true of hidden items and drop-down menus. |
| Optimize the number of rows in your URL's | <p>To use as few queries as possible, it's about having as many rows as possible on the website. Conversely, there is often an upper limit to what a web page will show, as too many rows can cause the scraper to go black.</p> <p>It is about finding the right level.</p> |
| Select one post at a time | Is it particularly complicated structures on a web page, then find links to the individual pages and scrape only one row at a time. |

Exercise

| What | How |
|--------------------------------|---|
| Get all tax depts in Wisconsin | <p>Wisconsin informs on tax depts on this website: https://www.revenue.wi.gov/Pages/Delqlist/NMALLB.aspx Not every depts is on the list. If you need you can see the regulations here: https://www.revenue.wi.gov/Pages/HTML/delqlist.aspx</p> <p>The link can't be read in import.io, because the table is an iframe. The direct link you can get by clicking on source code and search for iframe. For the table on A, the right link is: https://ww2.revenue.wi.gov/Internet/delqlist/NMALLA.htm</p> <p>By clicking on the different letters, you can uncover the link-structure:</p>  <p>For B it is https://ww2.revenue.wi.gov/Internet/delqlist/NMALLB.htm Only the letter A and B changes. The same principle is working for other letters:</p> <p>For # it is a bit different, namely: https://ww2.revenue.wi.gov/Internet/delqlist/NUMALL.htm</p> |
| Make Extractor | Place the link for the letter A in the box for New Extractor. Import.io should then suggest the scraper |
| Build all URL's | <p>Copy the letters from the website and paste them in Word:</p> <p><u>B</u> <u>C</u> <u>D</u> <u>E</u> <u>F</u> <u>G</u> <u>H</u> <u>I</u> <u>J</u> <u>K</u> <u>L</u> <u>M</u> <u>N</u> <u>O</u> <u>P</u> <u>Q</u> <u>R</u> <u>S</u> <u>T</u> <u>U</u> <u>V</u> <u>W</u> <u>X</u> <u>Y</u> <u>Z</u></p> <p>You can then run search and replace in Word on all spaces with ", ". Then you get.</p> <p><u>B</u>, <u>C</u>, <u>D</u>, <u>E</u>, <u>F</u>, <u>G</u>, <u>H</u>, <u>I</u>, <u>J</u>, <u>K</u>, <u>L</u>, <u>M</u>, <u>N</u>, <u>O</u>, <u>P</u>, <u>Q</u>, <u>R</u>, <u>S</u>, <u>T</u>, <u>U</u>, <u>V</u>, <u>W</u>, <u>X</u>, <u>Y</u>, <u>Z</u></p> <p>This can be used in URL-builder:</p>  |

Exercise

| | |
|---|---|
| | <p>Click Add to list. And add the latest url for # in the field:</p> <div><div>I</div><div>Enter or paste new url(s) here...</div></div> <p>Click Save. And check you don't have any url more than one time. Else delete.</p> |
| Run all URL's | <p>Click <u>RUN URL's</u>.</p> <div><div><div>Date/Time</div><div>11/22/18 18:25:39</div></div><div><div>Duration</div><div>00:00:30s</div></div><div><div>URLs</div><div>27</div></div><div><div>Success</div><div>27</div></div><div><div>Failed</div><div>0</div></div><div><div>Total Rows</div><div>18893</div></div></div> <p>In total 18.893 rows – no errors.</p> |
| Try perhaps also to make all url's in a spreadsheet | |

Exercise

| What | How |
|---|---|
| Examine Slovakian website | <p>Slovakia publish as other EU-members the farm subsidies on a website. You can find data for 2016 and 2017. http://www.apa.sk/en/information-about-beneficiaries-from-the-eagf-and-the-eafrd http://www.apa.sk/index.php?navID=202</p> <p>Chose 2017 and search. Look at the structure of the data at the site and how the URL's is build from page 1 to the last page. See if the first page also is shown if you use the same structure as the rest of the URL's</p> |
| Build New Extractor | Add the URL for the first page and check the result. Are you satisfied. You can't crawl down and see the detailed information. It need you to open another website. |
| Build URL-generator | After inspection of URL's, you can see the only variation is the page-number. Use that to build all the URL's in URL Builder. Remove Duplicates. |
| Run 3-4 websites to check | Don't run all the URL's. You only have 500 queries a Month and this scraper demands around the double of that limit. |
| Try the same with Croatian farmsubsidies | <p>Croatia publish also the farm subsidies on a website. You can find data for 2016 and 2017. http://isplate.apprrr.hr/ If you have time try the same as you did with Slovakian farm subsidies.</p> |
| Overview of EU farm subsidies | <p>The EU Commission has a website with links to all the different national websites for publishing farm subsidies: http://ec.europa.eu/agriculture/cap-funding/beneficiaries/shared_en</p> |

EXCERCISE

| What | How |
|--|---|
| Get a list of members | Open the link http://vl.dk/om-grupperne/gruppeoversigt/ You need the hidden link to iframe on the website with the overview. Use inspect. Search for iframe – or use inspect, net and record what is happening when you click. http://iframe.vl.dk/gruppeoversigt_new.php |
| Build an extractor on basis of the direct link to one group. | Klik på New Extractor og kopier det direkte link ind. Import.io analyserer data og kommer med et forslag. Import.io foreslår tre kolonner. Gem denne scraper ved at trykke Done. |
| Then build an extractor to extract the links to all groups. | |
| Use the second scraper as URL-input on the first scraper. | |