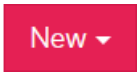
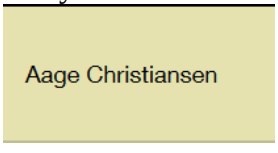
Kaas & Mulvad | 20.09.2016 | Data journalism | import.io

MEMO

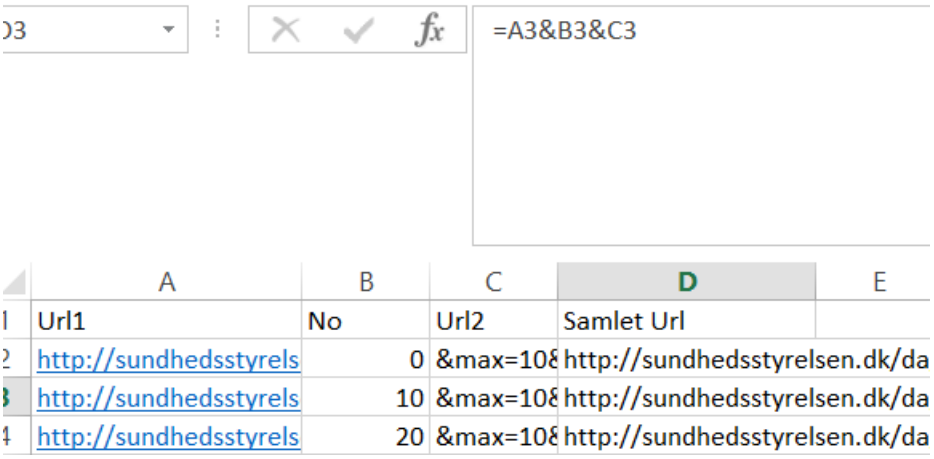
What	Why	How
Import.io scraper data without programming.	Import.io is a web scraper, which is automated to a high degree. Many tasks can be solved simply by copy and pasting the URL into its web page. You must register as a user to retrieve data as a csv file. It is free. You can also download a free version to your computer, where you earlier have been able to do more advanced scraping tasks. Finally, you can buy the enterprise version, and then import.io solves scraping tasks for you. All of their current and future product development efforts are now focused on the new web version of Import.io. They are working to bring across all of the old capability of the Desktop version, and they promise to include new features.	Basic site: https://import.io/ Free download: https://www.import.io/download/download-info/ Tutorial: http://support.import.io/ .
Magic	Magic is the most basic function, where you insert an url and then the software extracts the data in a structured way. If the data is on multiple pages, the limit is 20 pages. Magic handles a number of basic tasks - both in the web version and your downloaded version. import.io uses English separator of comma, which causes problems when loading other language data in Excel. However, it can be solved by using search and replace systematically (" ", " : and " , "to ;), moving columns and text columns where ";" is used as a separator. If you are importing a table from the web, it is often easier just to use an add-on for Google Chrome (Table Capture), which detects tables on a webpage, and you can then select the relevant table and then copy-paste into Excel. You can also use web import from Excel directly.	
Extractor	After installing import.io on your computer, you can use the more advanced methods. Extractor is not working on the web yet. Start by clicking New and select Extractor:  Insert the link into the browser window and press enter :  Change the button from OFF to ON:  Name the first column and the click on the first cell with the content of the column that you want extracted:  Click on Many rows.

MEMO

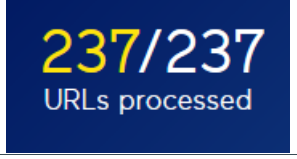
Give your number a header name. Click the + New column. Select the first cell in the next column. Give it a name until you have selected all the columns. Click Done.
Give your API a name and click Publish.
You can delete columns and you can later edit the Extractor.

Bulk Extraction

Once you have built your Extractor, you can put it to extract data from many URLs, if they have the same structure. You need to understand the method for going to the next page. The easiest way if it is just numbers from 1 to the last page like this:
http://www.fda.gov/Safety/Recalls/ArchiveRecalls/2015/default.htm?Page=1
in the last page:
http://www.fda.gov/Safety/Recalls/ArchiveRecalls/2015/default.htm?Page=15
Here you can make the 15 urls in a spreadsheet. First part of the url is this and it is constant:
http://www.fda.gov/Safety/Recalls/ArchiveRecalls/2015/default.htm?Page=
Second part is the number (1 to 15). You can merge that by the formula =A1&B1.
In total we will then build 15 URL's to be pasted into import.io
Here is a Google Spreadsheet with the URL's prepared: kortlink.dk/ks7f
In the example below the webpages are built by three parts.

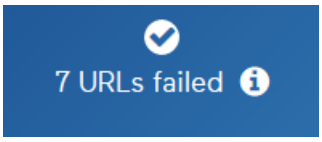


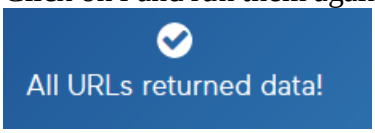
You collect the three parts of the formula: = A3 & B3 and C3 in cell D3. Then select all the 237 Url 's in Excel and copy them. You go back to import.io, choose Bulk Extract, insert the many urls and press Run Queries waiting. After a while, the screen shows that everything is pulled out :

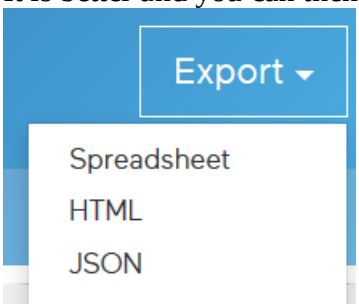


MEMO

<table>
<tr><td></td><td>
And yet, sometimes. Error detected :



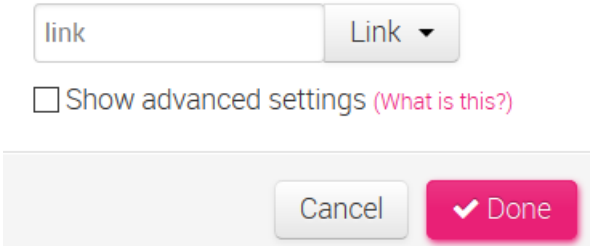
Click on i and run them again. Then it displays :


It is better and you can then download all - select spreadsheet


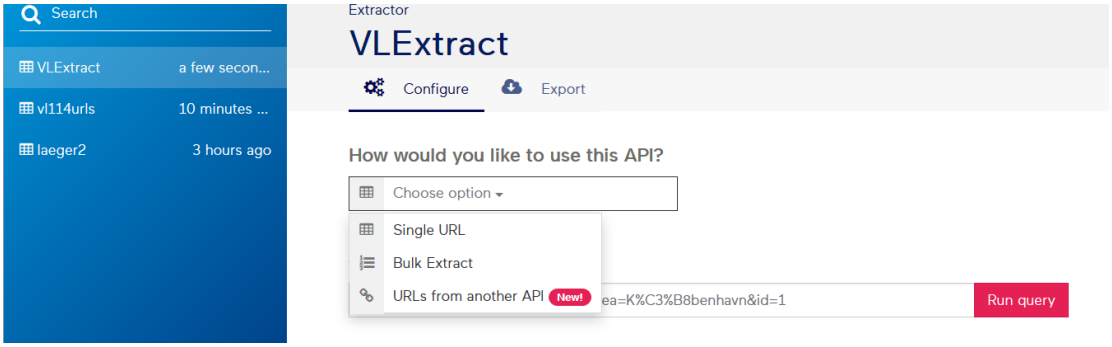
If you get problems with special characters, you can instead save it as HTML and open the file directly in Excel.
</td></tr>
<tr><td>Extraction of tax data</td><td>
Identify the structure of links in this page:
https://www.revenue.wi.gov/delqlist/nmalla.htm
For the letter A the url is:
https://www.revenue.wi.gov/delqlist/NMALLA.htm
For the letter B the url is:
https://www.revenue.wi.gov/delqlist/NMALLB.htm
And for the last url there is a slight change:
https://www.revenue.wi.gov/delqlist/NUMALL.htm
You can now construct the 27 urls copying the letters to word, replacing spaces with a new line, copy-pasting it to Excel and then make a formula combining 3 cells.
<table>
<tr><td>https://www.revenue.wi.gov/delqlist/NMALL</td><td>A</td><td>.htm</td><td>=a1&b1&c1</td></tr>
</table>

And then it’s the same as in the example below, making a new API from the first website and then copy and paste all the constructed urls into a bulk extraction.
</td></tr>
<tr><td>Extraction of many urls</td><td>
Often it is not so easy to identify the structure of links and you have to split the scraper in two API’s. One for extraction of links. The other for extraction of data. You then use the first API as input to the second.
</td></tr>
</table>

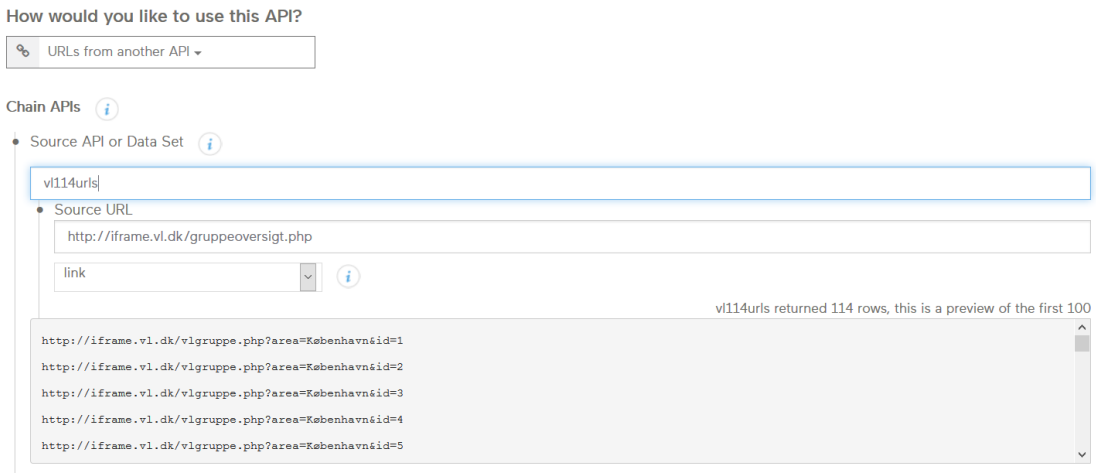
MEMO



You save this scraper and then build a new scraper to extract the data.



Then write the name of the relevant API into :



Plan for training in Kathmandu:	Sign up and get import.io downloaded. Good if it's done in advance. Then we will scrape these two websites: https://www.revenue.wi.gov/delqlist/nmalla.htm http://www.fda.gov/Safety/Recalls/ArchiveRecalls/2015/default.htm The first website we will try in Magic. And when that is done we will do a bulk extract from magic, copy-pasting the links from this Google Spreadsheet: https://docs.google.com/spreadsheets/d/1_T7eT3VvR2gPW8JFXX-RCnNEuyY7cCdUmFHYzShYpH8/edit?usp=sharing And then we will download the data as a spreadsheet and html and open it in Excel. Next step is to use the Extractor on the second website, the fda-data. And then also from the spreadsheet copy-pasting the links for bulk extraction. Run it and then download and clean the data. In the end a demo on how one API can be the input for another scraper.