

Lab 5: Data Scraping

To better understand how you can take advantage of the data scraping functionality, let's create an automation project that extracts some specific information from Wikipedia and writes it to an Excel spreadsheet. You can use this type of automation in different scenarios, such as extracting lists of products and their prices from e-commerce websites.

Note:

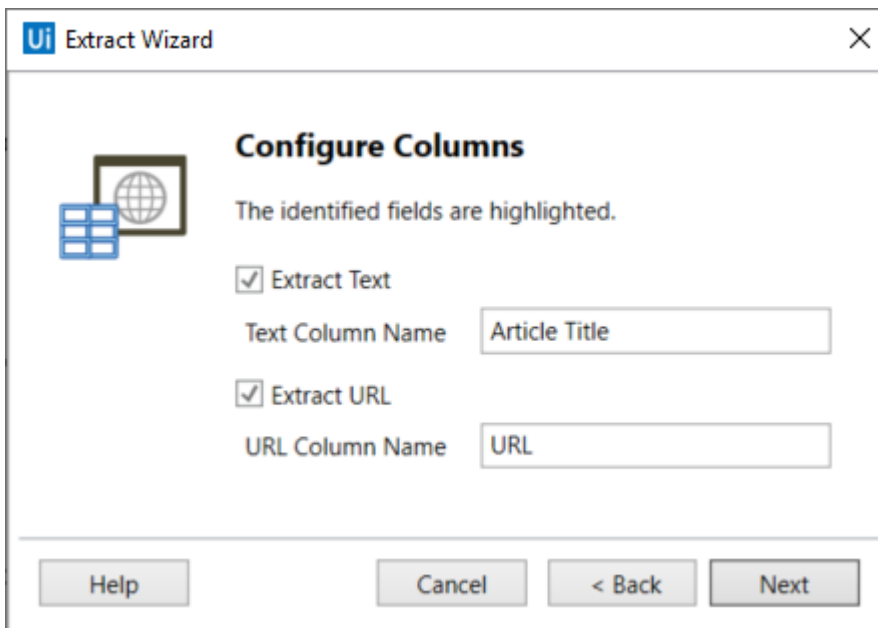
It is recommended to run your web automations on Internet Explorer 11 and above, Mozilla Firefox 50 or above, or the latest version of Google Chrome.

Lab Solutions

Lab solution(s) are present in `Solution\Lab05` folder.

Let's say you want to start reading up on economics and you want to get a list of Wikipedia articles on the subject, together with their URLs, and the additional information that is provided in the search results for each article. You can do the following:

1. Open Internet Explorer and navigate to en.wikipedia.org.
2. In the **Search Wikipedia** box, type "economics", then click "containing... *economics*" in the drop-down that appears. A web page opens displaying the search results.
3. In Studio, create a **New Blank Process**.
4. From the **Activities** panel, add an **[Open Browser]** activity to the **Designer** panel and, in the **Url** field, paste the URL of the web page with the search results. In our example, the URL is:
"<https://en.wikipedia.org/w/index.php?search=economics%20&title=Special%3ASearch&fulltext=1&ns0=1>".
5. In the **Design** ribbon tab, in the **Wizards** group, click **Data Scraping**. The **Extract Wizard** is displayed.
6. Following the wizard, select the first and last items on the web page. The **Configure Columns** wizard step is displayed and the fields you selected are highlighted in the web browser.
7. Select the **Extract URL** check box and change the name of the column headers to something relevant, for example "Article Title" and "URL".



Ui Extract Wizard

Configure Columns

The identified fields are highlighted.

☒ Extract Text

Text Column Name

☒ Extract URL

URL Column Name

8. Click **Next**. A preview of the data is displayed in the **Preview Data** wizard step. Note that because the Wikipedia page uses relative URLs, the URL column contains relative URLs as well. You can correct this in the Excel output after the project is executed by adding the string "<https://en.wikipedia.org>" at the beginning of each cell in the URL column.

The screenshot shows the 'Extract Wizard' window with the 'Preview Data' step. The background is a Wikipedia search results page for 'economics'. The wizard window has a table titled 'Preview Data' with two columns: 'Article Title' and 'URL'. The table lists 20 articles related to economics, such as 'Economics', 'Keynesian economics', 'Positive economics', etc. The 'URL' column contains relative paths like '/wiki/Economics'. At the bottom of the wizard, there are buttons for 'Help', 'Cancel', '< Back', 'Extract Correlated Data' (highlighted with a red box and an arrow), and 'Finish'. There is also a text input field for 'Maximum number of results (0 for all)' set to '100'.

Article Title	URL
Economics	/wiki/Economics
Keynesian economics	/wiki/Keynesian_economics
Positive economics	/wiki/Positive_economics
Environmental economics	/wiki/Environmental_economics
Neoclassical economics	/wiki/Neoclassical_economics
Heterodox economics	/wiki/Heterodox_economics
Socialist economics	/wiki/Socialist_economics
Agricultural economics	/wiki/Agricultural_economics
Development economics	/wiki/Development_economics
Labour economics	/wiki/Labour_economics
Behavioral economics	/wiki/Behavioral_economics
Master of Economics	/wiki/Master_of_Economics
Welfare economics	/wiki/Welfare_economics
Classical economics	/wiki/Classical_economics
Capital Economics	/wiki/Capital_Economics
Bachelor of Economics	/wiki/Bachelor_of_Economics
Microeconomics	/wiki/Microeconomics
Gross (economics)	/wiki/Gross_(economics)
Health economics	/wiki/Health_economics
Managerial economics	/wiki/Managerial_economics

9. Click the **Extract Correlated Data** button to extract additional information about the articles. The **Extract Wizard** starts again.
10. Following the wizard again, indicate the information about the size and date of the last edit that is available for each article. The **Configure Columns** step is reached again.
11. Change the name of the new column header to "Additional Information" and click **Next**. The data is displayed in the **Preview Data** wizard step. Optionally, you can change the order of the columns by dragging them in place.
12. In the **Maximum number of results** type 60. The Wikipedia search lists 20 results per page and, for our example, we want to extract the first three pages of search results.

Ui

Extract Wizard

×

Preview Data

Article Title	URL	Additional Info
Economics	/wiki/Economics	167 KB (18,415 words) - 14:10, 22 N
Keynesian economics	/wiki/Keynesian_economics	83 KB (10,775 words) - 05:41, 10 M
Positive economics	/wiki/Positive_economics	8 KB (838 words) - 16:17, 19 May 21
Environmental economics	/wiki/Environmental_economics	28 KB (3,400 words) - 13:47, 22 Ma
Neoclassical economics	/wiki/Neoclassical_economics	25 KB (3,209 words) - 21:43, 17 Ma
Heterodox economics	/wiki/Heterodox_economics	28 KB (2,906 words) - 06:15, 22 Ma
Socialist economics	/wiki/Socialist_economics	131 KB (17,468 words) - 18:23, 24 A
Agricultural economics	/wiki/Agricultural_economics	17 KB (1,825 words) - 00:03, 15 Ma
Development economics	/wiki/Development_economics	52 KB (6,160 words) - 13:46, 19 Ma
Labour economics	/wiki/Labour_economics	36 KB (4,588 words) - 22:06, 23 Apr
Behavioral economics	/wiki/Behavioral_economics	80 KB (8,427 words) - 18:48, 21 Ma
Master of Economics	/wiki/Master_of_Economics	5 KB (534 words) - 16:02, 22 July 20
Welfare economics	/wiki/Welfare_economics	23 KB (2,793 words) - 05:16, 21 Ma
Classical economics	/wiki/Classical_economics	20 KB (2,542 words) - 18:24, 11 Ma
Capital Economics	/wiki/Capital_Economics	3 KB (251 words) - 11:57, 19 Noven
Bachelor of Economics	/wiki/Bachelor_of_Economics	3 KB (271 words) - 19:18, 6 March 2
Microeconomics	/wiki/Microeconomics	32 KB (3,586 words) - 05:34, 20 Ma
Gross (economics)	/wiki/Gross_(economics)	430 bytes (7 words) - 15:39, 26 Feb
Health economics	/wiki/Health_economics	20 KB (2,419 words) - 21:31, 14 Apr
Managerial economics	/wiki/Managerial_economics	9 KB (800 words) - 09:08, 4 May 20

Edit Data Definition

Maximum number of results (0 for all)

60

Help

Cancel

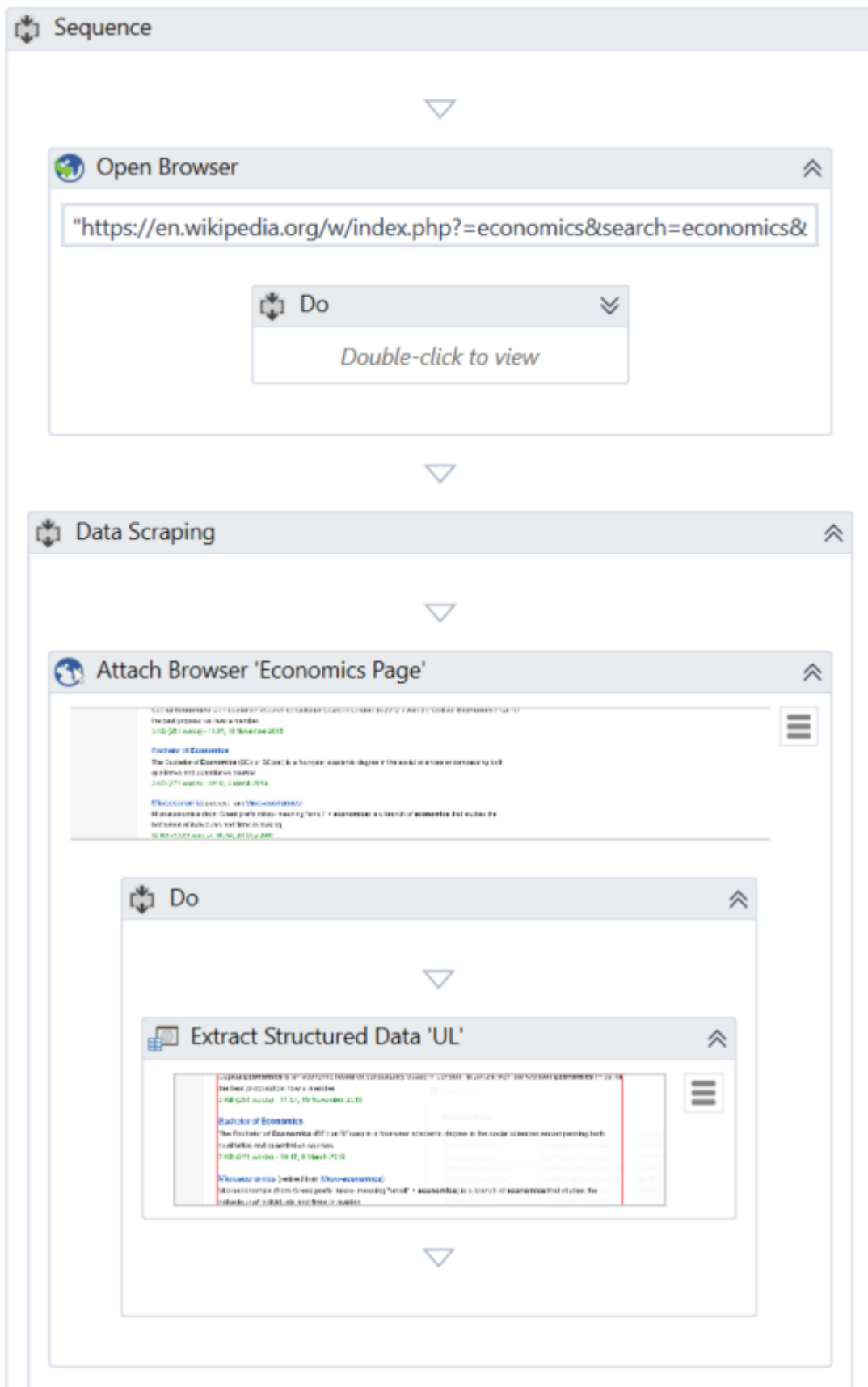
< Back

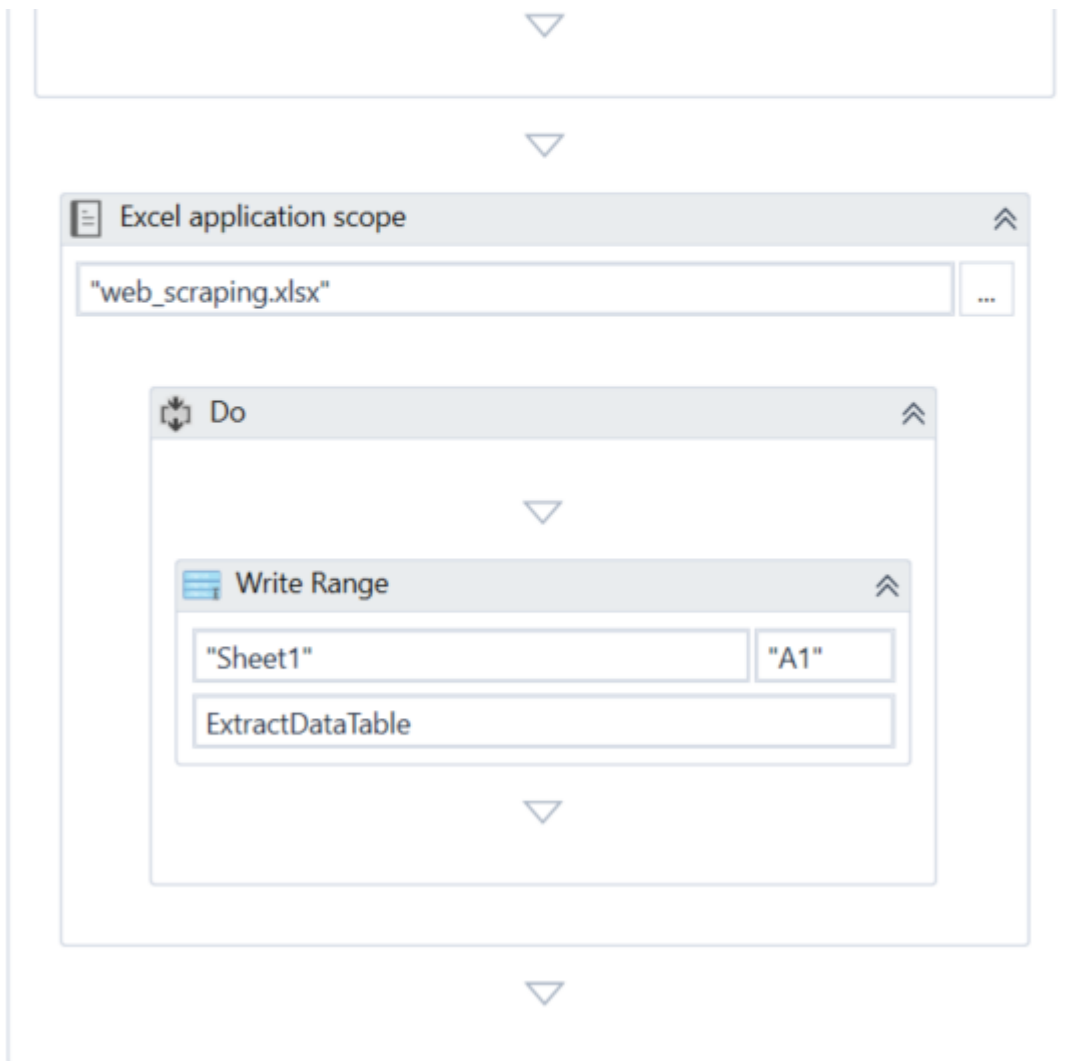
Extract Correlated Data

Finish

- Click **Finish**. The **Indicate Next Link** window is displayed prompting you to indicate the **Next** button or arrow to click if the data spans more than one page.
- Click **Yes** and select the **next 20** button below the search results in Wikipedia. The project is updated and a **Data Scraping** sequence is displayed in the **Designer** panel. A **DataTable** variable, ExtractDataTable has been automatically generated.
- In the **Variables** panel, change the scope of the automatically generated ExtractDataTable variable to **Sequence**. Do this to make the variable available outside of its current scope, the **Data Scraping** sequence.
- Add an **[Excel Application Scope]** activity under the **Data Scraping** sequence.
- In the **Properties** panel of the Excel Application Scope activity, in the **WorkbookPath** field, type "web_scraping.xlsx". Upon project execution, a file with this name is created in the project folder to store data from the scraping. Alternatively, you can specify a file that already exists on your machine.
- In the **Do** sequence of the **Excel Application Scope** activity, add a **[Write Range]** activity and in the **Properties** panel:
 - In the **DataTable** field, add the ExtractDataTable variable.
 - Select the **AddHeaders** check box to include the column names in the output.

The final project should look as in the following screenshot:





19. Press **F5** to execute the project.

20. Open the Excel file you defined in step 17. Note that all columns are populated correctly.

File Home Insert Page Layout Formulas Data Review View Help Search		
<div> <div>Clipboard</div> <div>Font</div> <div>Alignment</div> <div>Number</div> <div>Styles</div> <div>Conditional Formatting</div> <div>Format as Table</div> <div>Cell Styles</div> <div>Cells</div> <div>Editing</div> <div>Ideas</div> </div>		
D1		
A	B	C
1 Article Title	URL	Additional Info
2 Economics	/wiki/Economics	167 KB (18,415 words) - 14:10, 22 May 2019
3 Keynesian economics	/wiki/Keynesian_economics	83 KB (10,775 words) - 05:41, 10 May 2019
4 Positive economics	/wiki/Positive_economics	8 KB (838 words) - 16:17, 19 May 2019
5 Environmental economics	/wiki/Environmental_economics	28 KB (3,400 words) - 13:47, 22 May 2019
6 Neoclassical economics	/wiki/Neoclassical_economics	25 KB (3,209 words) - 21:43, 17 May 2019
7 Heterodox economics	/wiki/Heterodox_economics	28 KB (2,906 words) - 06:15, 22 May 2019
8 Socialist economics	/wiki/Socialist_economics	131 KB (17,468 words) - 18:23, 24 April 2019
9 Agricultural economics	/wiki/Agricultural_economics	17 KB (1,825 words) - 00:03, 15 May 2019
10 Development economics	/wiki/Development_economics	52 KB (6,160 words) - 13:46, 19 May 2019
11 Labour economics	/wiki/Labour_economics	36 KB (4,588 words) - 22:06, 23 April 2019
12 Behavioral economics	/wiki/Behavioral_economics	80 KB (8,427 words) - 18:48, 21 May 2019
13 Master of Economics	/wiki/Master_of_Economics	5 KB (534 words) - 16:02, 22 July 2018
14 Welfare economics	/wiki/Welfare_economics	23 KB (2,793 words) - 05:16, 21 May 2019
15 Classical economics	/wiki/Classical_economics	20 KB (2,542 words) - 18:24, 11 March 2019
16 Capital Economics	/wiki/Capital_Economics	3 KB (251 words) - 11:57, 19 November 2018
17 Bachelor of Economics	/wiki/Bachelor_of_Economics	3 KB (271 words) - 19:18, 6 March 2019
18 Microeconomics	/wiki/Microeconomics	32 KB (3,586 words) - 05:34, 20 May 2019
19 Gross (economics)	/wiki/Gross_(economics)	430 bytes (7 words) - 15:39, 26 February 2019
20 Health economics	/wiki/Health_economics	20 KB (2,419 words) - 21:31, 14 April 2019
21 Managerial economics	/wiki/Managerial_economics	9 KB (800 words) - 09:08, 4 May 2019
22 Positive economics	/wiki/Positive_economics	8 KB (838 words) - 16:17, 19 May 2019
23 Home economics	/wiki/Home_economics	30 KB (3,551 words) - 21:18, 13 May 2019
24 Managerial economics	/wiki/Managerial_economics	9 KB (800 words) - 09:08, 4 May 2019
25 Institutional economics	/wiki/Institutional_economics	28 KB (3,087 words) - 09:00, 17 May 2019
26 Trickle-down economics	/wiki/Trickle-down_economics	18 KB (2,034 words) - 10:51, 20 May 2019