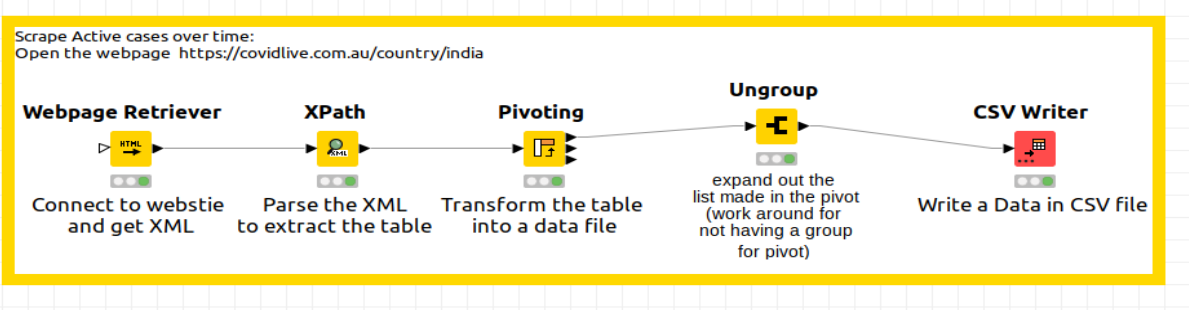
Web Scraping with KNIME

Introduction:

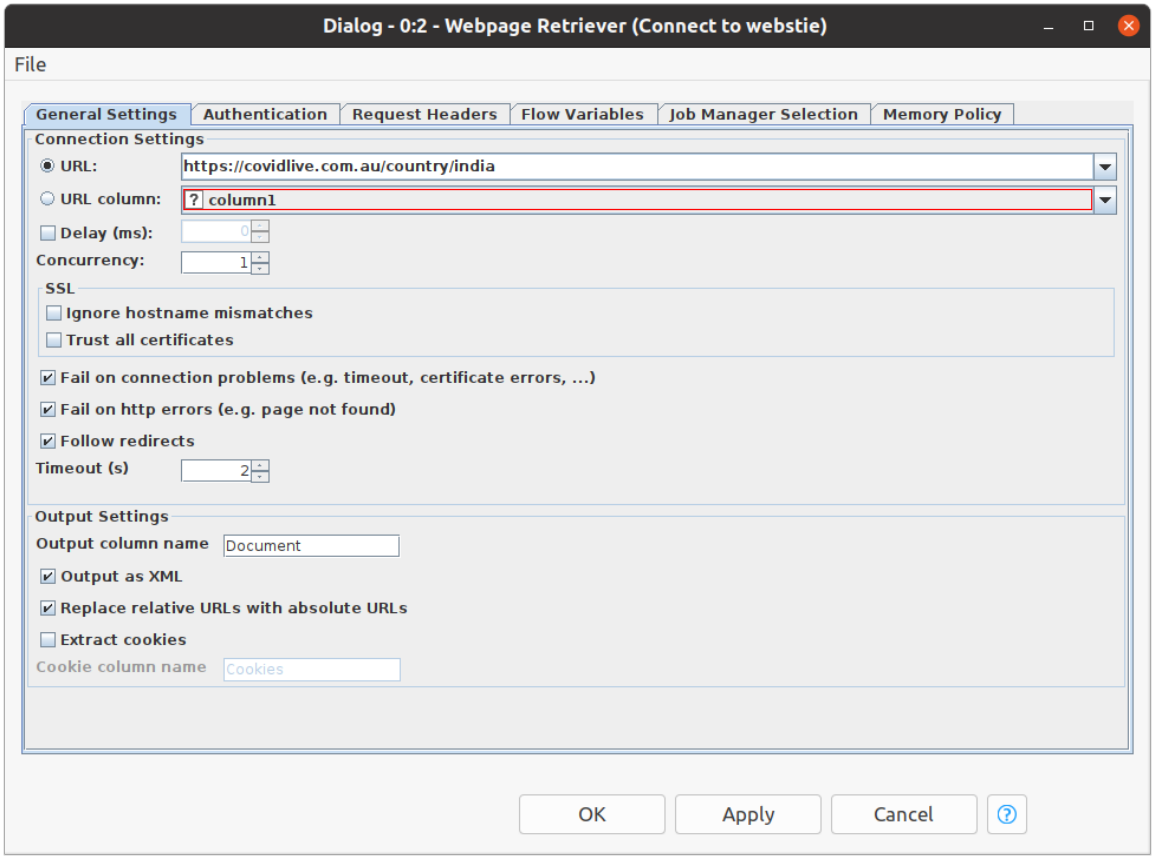
Web scraping is the most important thing in Data Science. Our objective in this article is to make a simple workflow in KNIME to see How KNIME scraps the data from the website. Here we have to take a live website of corona cases and other information for India, you can find the website here.



The above image is the final workflow of web scraping. We are going step by step and understand how we can configure this node and fetch the data from the website. Before we start you must have some of the basics knowledge about the KNIME Analytics platform and you can refer to the previous blog here.

1) Webpage Retriever Node

- Webpage Retriever Node provides the facility to connect with the website and generate the XML file, so as you can see in the below dialog box of wen retriever node you have to just specify your respective URL in the connection setting in our case we have to specify (26,752,447 Cases in India - COVID LIVE) this web site in URL.
- You can find a more detailed description of the webpage Retriever Node from here.

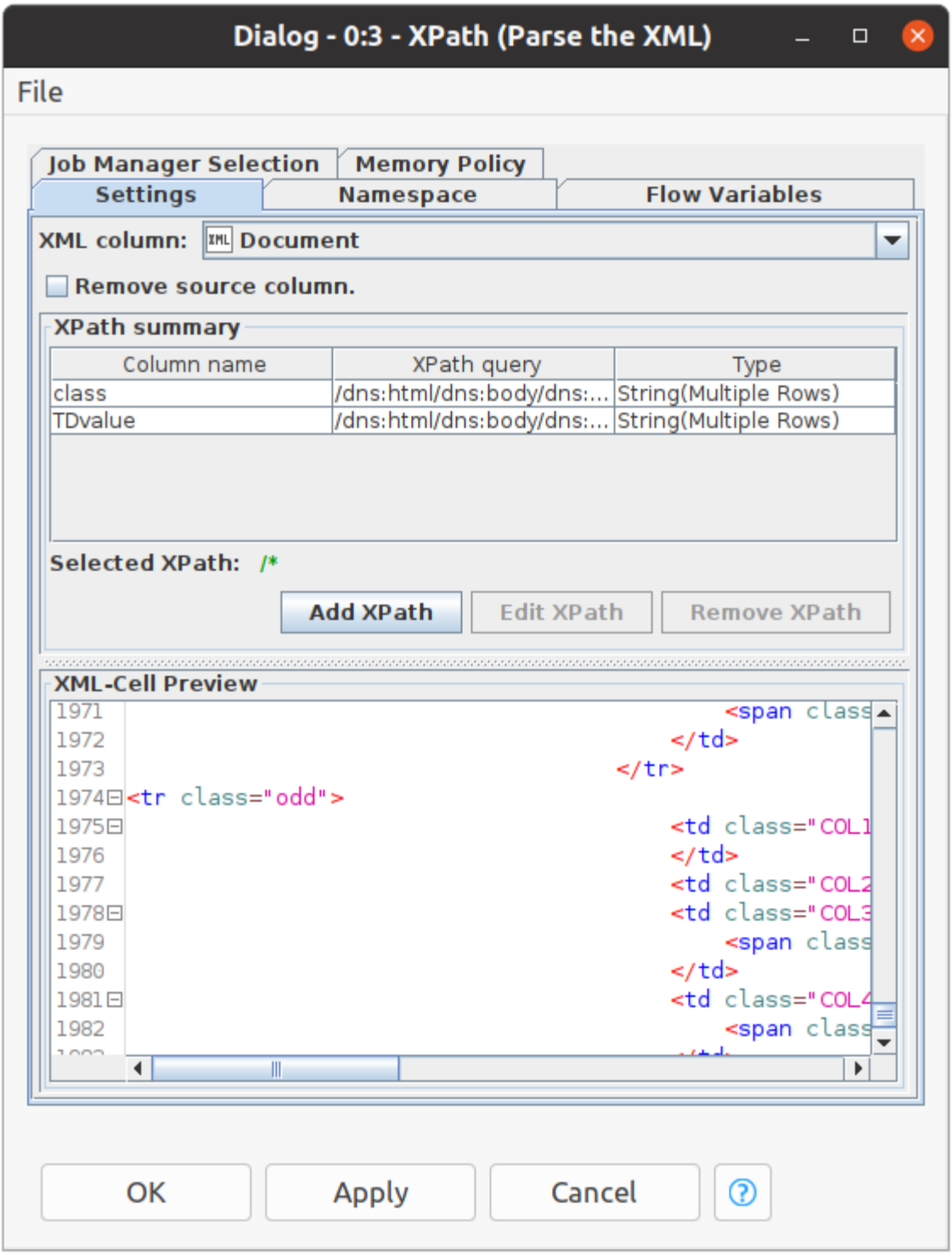


.

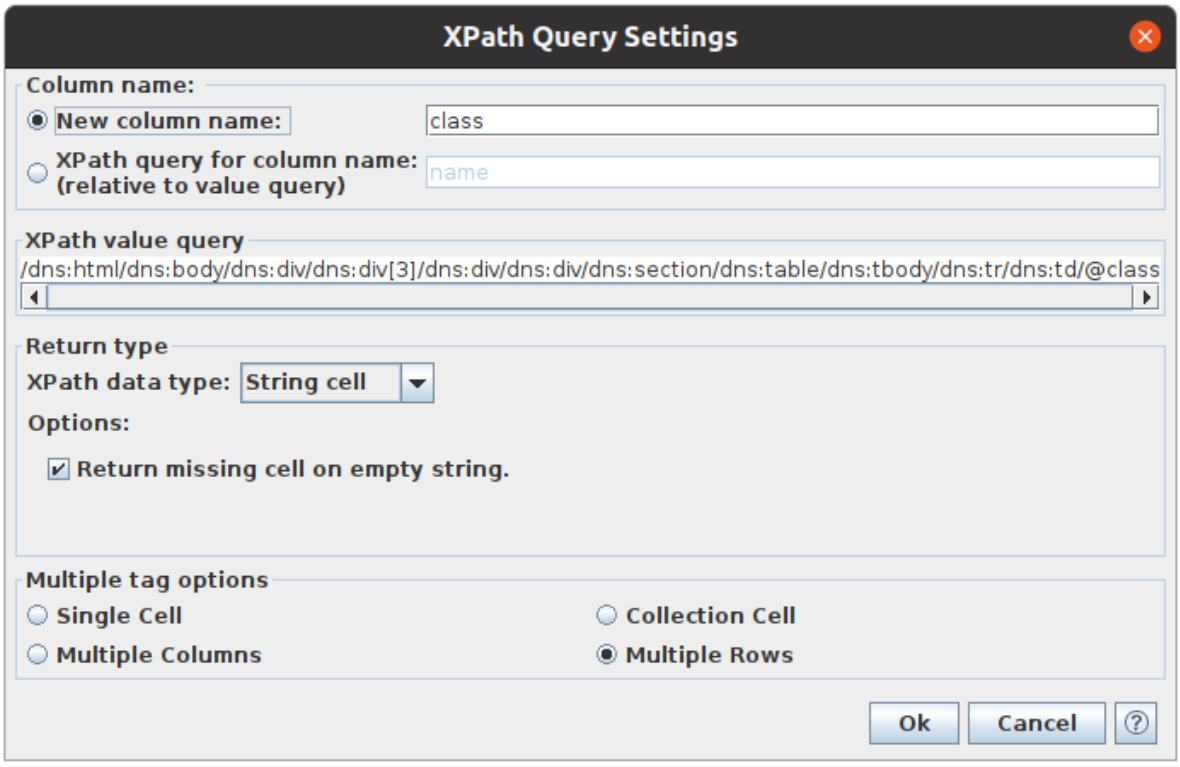
Press OK and run the Node and you are getting the output of this URL in XML format

2) XPath

- So, our next step is to parse the XML file to extract the table data from it. **XPath Node** helps you out to parse the XML file, as you can see below the configuration of Xpath we have to specify the Xpath.
- Using the Add path button you can add a different Path and XPath Summary displays how many paths you have.
- You can find more information about XPath from here to get more ideas about how XPath works.

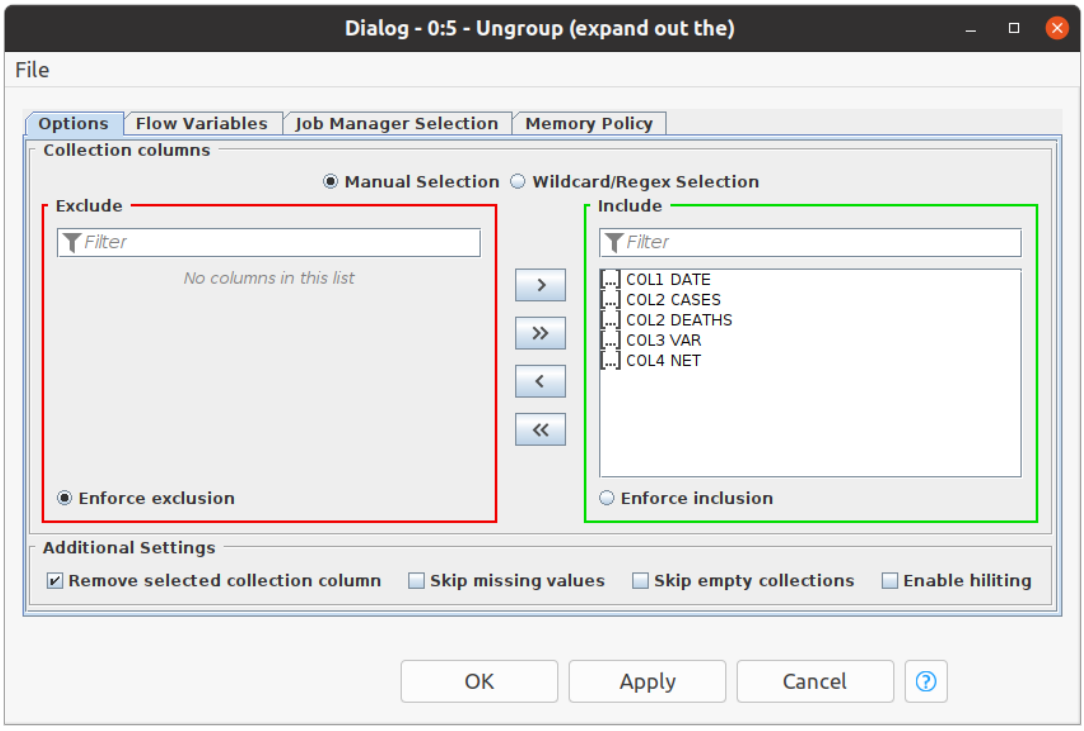


- Now, We have to write an XPath query to select the attributes and class, etc to fetch the data, as you can see in the below dialog of XPath Query You have to specify s the column name and write a query in **XPath value query**.
- If you are new to XML and XPath query you can refer to this link and get an idea of the syntax of the XPath query.



3) Pivoting and Ungroup

- Now, our next step is to transform the table into a data file we use Pivoting Node. And after this, we have to Ungroup the column using **Ungroup Node**.



4) CSV Writer

- Our final step is to write a CSV file of this fetched data and KNIME Provides facilities to do this by simply using CSV writer Node.
- So, By executing this full workflow at the end in the CSV you get the fetch data of date-wise corona cases and deaths in India.

Authored by:

Team Cilans: *Nikhil, Chintan, Kashyap*

For additional articles on the **Knime Blog series** visit www.cilans.net

Contact us at info@cilans.net for any queries.