Notes : June 16, 2021

Project : **[Download files, Extract Data and Store data](https://www.freelancer.com/projects/web-scraping/Download-files-Extract-Data-Store" \t "_self)**

Jeff Jones : I just sent you a new set of files with two new columns (Report Date and Ticker).

Brian Bartock : The NLY iXBRL file contains 2,144 Facts, your CSV file has 2,071. Any idea where the difference is coming from?

Jeff Jones : I filtered some of the text fields out, as I thought you were focusing on leverage calculations-. I have removed those filters and the Facts number in the XBRL viewer files now matches the CSV files.

Brian Bartock : In some of the files they have a format field to designate $millions, can that added be added as column? For example in the NLY file, CashAndCashEquivalentsAtCarryingValue is 222 but shows up in the text as 22.2.

Jeff Jones : The CashAndCashEquivalentsAtCarryingValue is stated as 22.2 in a footnote where it is described as being in millions. This is an example of where I think there might be some issues with the XBRL viewer or perhaps I am missing something. The value in the XML file is 22200000 and the ‘decimals’ field is -5. This would result in a value of 222 (which is what is in the produced CSV). However, in the footnote on the report it states CashAndCashEquivalentsAtCarryingValue 22.2 as being in millions. The ‘decimals’ value in the XML file should be -6.

Brian Bartock : Do you mind adding the ticker as a column?

Jeff Jones : Not a problem. This is in the current latest set of files I have sent you.

Brian Bartock : As I start merging some datasets it will make my life easier. In the instance where we go back a quarter, how would you handle not restoring a lot of the same values?

Jeff Jones : If you could explain what you are referring to, that would help. If a field is published on a report on a certain date, the report date and the ‘as of’ date of the field would be used as eys to store it (and subsequently retrieve it. So, if the CashAndCashEquivalentsAtCarryingValue for 12/31/2020 is published on the 12/31/2020 and the 3/31/2021 10Q’s, they are stored as separate pieces of data (so you can see if they have restated it). If that isn’t what you need, happy to change it (right now, it is just writing to the CSV file so no change really required.

Brian Bartock : In the AAIC file, I see a few values of “Suppressed”. Any idea what that means?

Jeff Jones : There are some text fields which include a lot of additional formatting information like fonts and colors. I took the liberty of suppressing them as they have special characters which are not part of the standard ASCII character set (so you can’t print them or view them in a CSV file). We should discuss if you need those.

Brian Bartock : When going back to prior quarters, will the supplemental Schema and presentation/calculation files also need to be downloaded?

Jeff Jones : I did not download them as I thought everything you needed was in the main XBRL instance (xml file). However, if the schema’s changed drastically in some of the older 10Q’s, I might have to.

Brian Bartock : Looking at some of these tickers, some of them didn’t file the full XML package going back to 2010. If you ran one of the old .htm files that have the xml tags, would it handle it?

Jeff Jones : I can try that tomorrow. I will let you know.

Brian Bartock : If not, I’m fine going back as far as the XML packages are provided. As a next step, can we grab one particular company and pull the historical Qs and scrape them?

Jeff Jones : You seem focused on NLY. I will work on that one unless you have some others you would rather have me look at.