Wikipedia Physicists

Tim Evans

6th January 2022

# How to download biography on each physicist from Wikipedia

First note that there is a real problem with non-ASCII characters in names, especially on Windows. Windows does NOT use UTF8 encoding (unlike Linux and MacOS) so it seems that when python is printing to screen and writing to text files on Windows it may garble these characters. On Windows most applications do not do this, e.g. text editors and Excel. There is an option in python 3.7 to force UTF8 encoding on Windows so I read but I have not tested this.

I would always work with a shortest list when testing and include a few names with non-ascii characters e.g. "André-Marie Ampère" and "Anders Jonas Ångström".

In practice some of the editing is done quickest by hand in an application like Excel.

I started from

[List of physicists - Wikipedia](https://en.wikipedia.org/wiki/List_of_physicists)

<https://en.wikipedia.org/wiki/List_of_physicists>

This can be saved and the irrelevant information removed. A simplified list can be found in my file

ListPhysicistsWikipedia.html

I then used excel to rearrange this by hand into

Wikipedia-20220106192724-ListOfPhysicists-NamesInfo.xlsx

I recommend that you down load a copy of all the pages in some formal just ONCE. You note the time and day and then this becomes your core data. The pages do change so by sticking with one local data set your results are reproducible.

## Using Bulk Download on Wikipedia

There are several ways to get a bulk download, including the whole of Wikipedia. This may be too large to work with on a personal machine.

One simpler route is to use the Special:Export page. Goto

<https://en.wikipedia.org/wiki/Special:Export>

In the "Add pages manually:" box you need the "title" of each physicist page with accents such as "André-Marie Ampère" so "Andr%C3%A9-Marie\_Amp%C3%A8re" does not work

Use list from first column of Wikipedia-20220106192724-ListOfPhysicists-NamesInfo.xlsx (perhaps in batches of around 500) to paste into "Add pages manually:" box. This gives you an xml file which can be read in python. Take a look at the start and end.

By the way, in the "Add pages from category" if you enter "physics" in box and you will get a list of known physics related pages such as "Category:Physicists" in the "Add pages manually:" box. However I have not worked out how to use this yet.

I recommend you do this in full just ONCE. You note the time and day and then this becomes your core data. The pages do change so this way your results are reproducible.

## Using Direct Access (scraping) from Python

I have another route which loads each page direct from python but this is really not good for the servers and Wikipedia asks you not to do this. I would suggest you use this only if needed and then to download pages ONCE to disk and then you use your local files after that. I have some code working as again UTF8 and recursion problems were severe and it took several attempts to get this to work.

A better way and one favoured by Wikipedia is to use the API which is what my student Henry Price (Henry.Price10@imperial.ac.uk) did but he only got a partial download (not sure what happened there).