There is a total of 62,181 pages in the Lamoni Chronicle Achieve. Because you can extract the GETpdf URL from the <https://lamoni.advantage-preservation.com> website you don’t have to store the pdfs on a drive you can just have the Docling converter pipeline access the GETpdf URL.

**Files**

UrlSaver.py gets the GETpdf urls and names of each pdf from the Lamoni preservation website. It then saves those pieces of information in GetPDFUrls.csv in two columns. Every GETpdf url takes around 1.23 seconds to save. For the entire Lamoni archive this would take around 21 hours give or take depending on the response time of the Lamoni archive.

CSVDigitalizer.py extracts the GETpdf and name from each row int he the GetPDFUrls.csv file. Then using the Docling library to digitalizes the newspaper pdf using the GETpdf url. Finally, it names the scan with the corresponding name from GetPDFUrls.csv and saves it to the RawDigitalizedScans directory. Each file takes on average 94 seconds to digitalize on my Mac but thankfully Docling can leverage GPUs to make this digitalization go faster.

DirectWebDegitalizer.py Does what the above two files do but in one python file and without using an intermediate csv file to store the GETpdf urls and names.

Bugs

For some reason the UrlSaver.py file will execute for a while and then pause for some reason. The code doesn’t crash or anything which is frustrating. Might have something to do with my laptop that I’m running it on or worst case the actual website itself having a hiccup in its API.

UrlSaver had an error at url 49544:

File "/Users/jackflickinger/Desktop/GracelandGPT/UrlSaver.py", line 125, in <module>

fileName = CreateFileName(url)

^^^^^^^^^^^^^^^^^^^

File "/Users/jackflickinger/Desktop/GracelandGPT/UrlSaver.py", line 45, in CreateFileName

fileName = raw\_iframe.split(start)[1].split(end)[0]

~~~~~~~~~~~~~~~~~~~~~~~^^^

IndexError: list index out of range