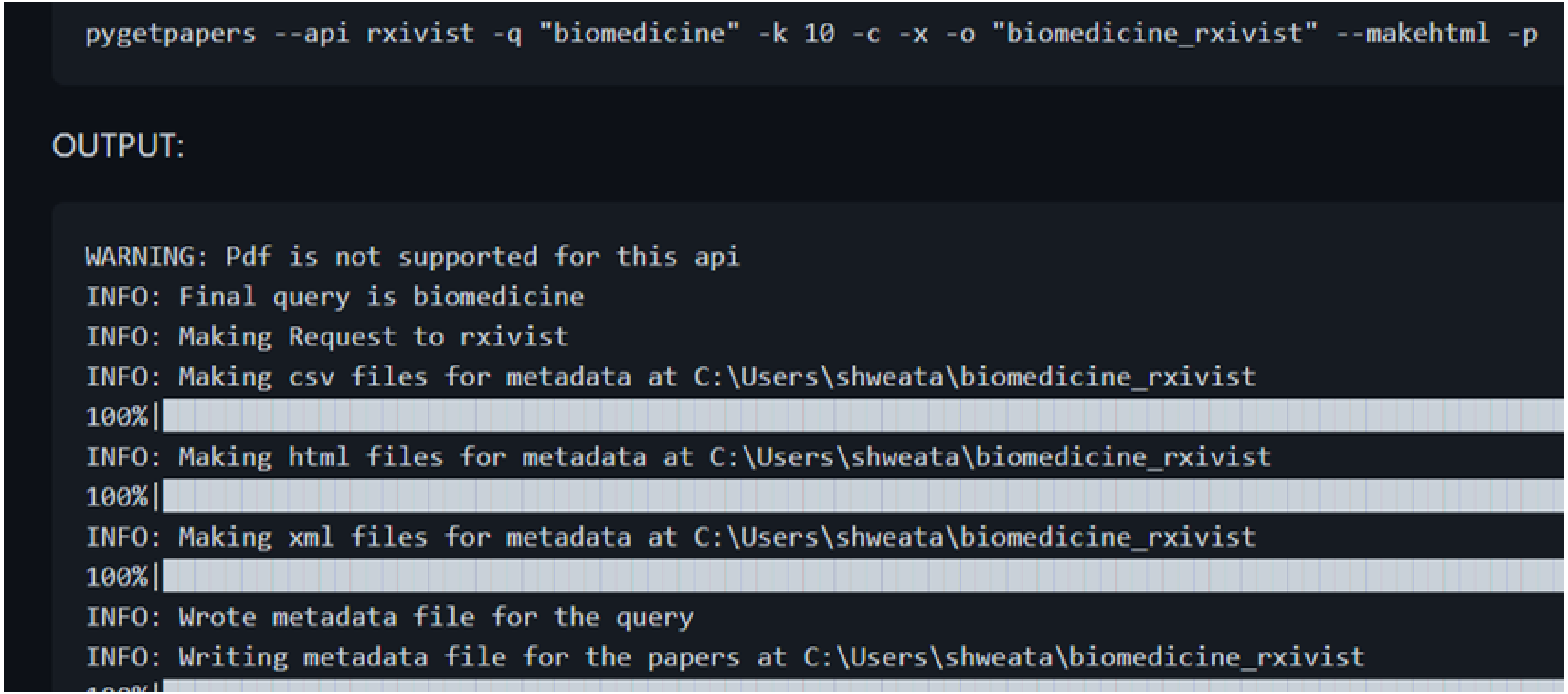
pygetpapers

PYGETPAPERS

A PYTHON REWRITE OF RICK SMITH-UNNA'S GETPAPERS BY AYUSH GARG

pygetpapers is a tool to assist text miners. It makes requests to open access scientific text repositories, analyses the hits, and systematically downloads the articles without further interaction.



```
pygetpapers --api rxivist -q "biomedicine" -k 10 -c -x -o "biomedicine_rxivist" --makehtml -p
```

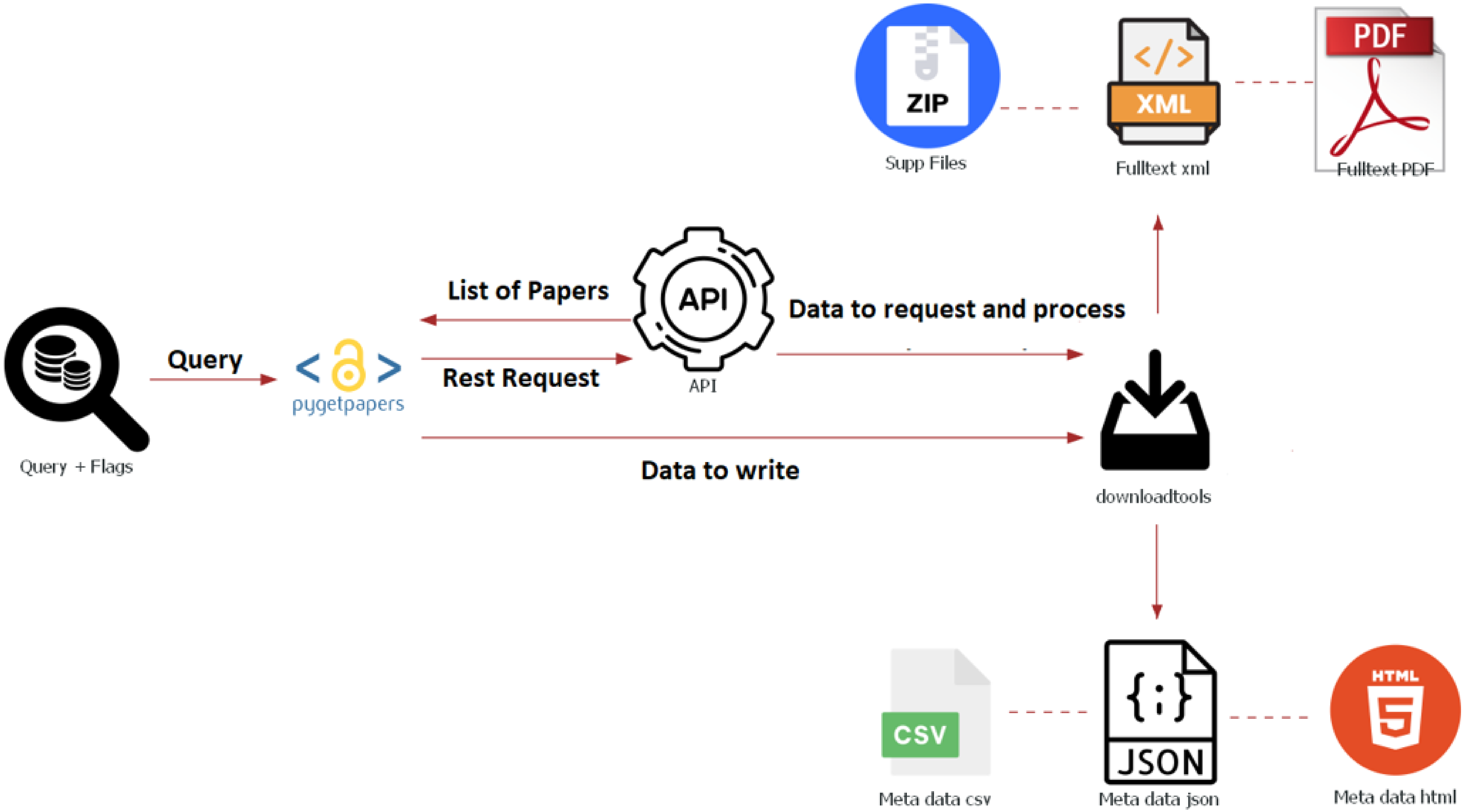
OUTPUT:

```
WARNING: Pdf is not supported for this api
INFO: Final query is biomedicine
INFO: Making Request to rxivist
INFO: Making csv files for metadata at C:\Users\shweata\biomedicine_rxivist
100%|
INFO: Making html files for metadata at C:\Users\shweata\biomedicine_rxivist
100%|
INFO: Making xml files for metadata at C:\Users\shweata\biomedicine_rxivist
100%|
INFO: Wrote metadata file for the query
INFO: Writing metadata file for the papers at C:\Users\shweata\biomedicine_rxivist
```

TECHNICAL DESIGN

Supplemental

Full text



Meta Data

STEP BY STEP USAGE OF PYGETPAPERS

1)Create Query

2)Choose the repository

3)Download full text/metadata/supplemental

→

PYGETPAPERS MAKES THIS PROCESS **AUTOMATED**

QUERY BUILDER

Lets say we have to query for Lantana Camara

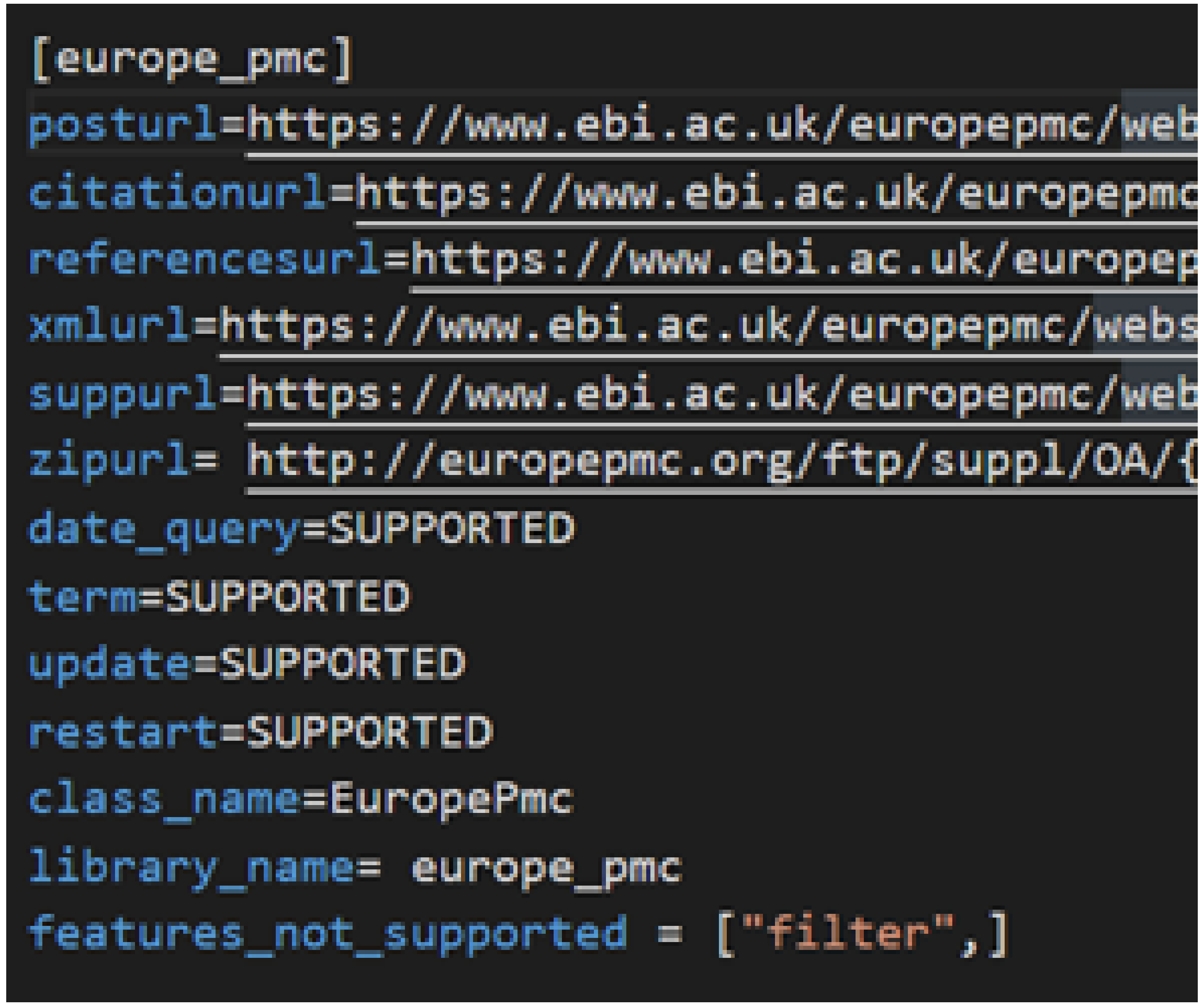
1)Search within a date range (--startdate , --enddate)
2)Build query with ami dictionaries (--term, --notterms)
3)Build query with terms in a text file (--term, --notterms)
4)Compound Queries (AND, OR, NOT)

API SUPPORT

Config File

Currently Supported APIs:

- Europe PMC
- arXiv
- bioRxiv
- medRxiv
- crossref
- Rxivist



```
[europe_pmc]
posturl=https://www.ebi.ac.uk/europepmc/web
citationurl=https://www.ebi.ac.uk/europepmc
referencesurl=https://www.ebi.ac.uk/europep
xmlurl=https://www.ebi.ac.uk/europepmc/webs
suppurl=https://www.ebi.ac.uk/europepmc/web
zipurl= http://europepmc.org/ftp/suppl/OA/{
date_query=SUPPORTED
term=SUPPORTED
update=SUPPORTED
restart=SUPPORTED
class_name=EuropePmc
library_name= europe_pmc
features_not_supported = ["filter",]
```

pygetpapers uses a plugin based approach so adding new repositories is super easy

KEY FEATURES OF PYGETPAPERS

1)Updating corpus with new papers

2)Adding new types of metadata/data to the corpus

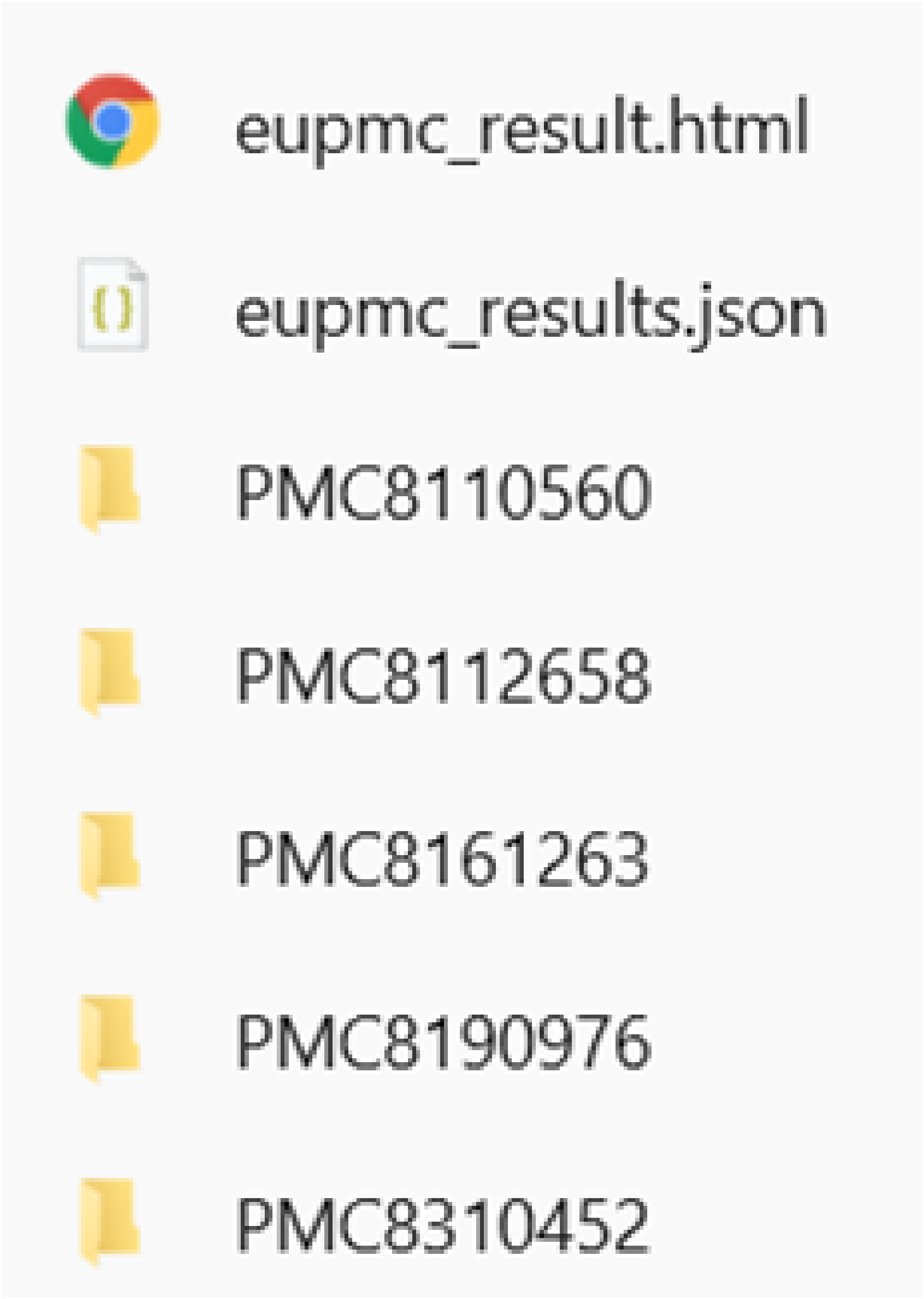
3)Changing the log levels

4)Saving logs

5)Saving query to reuse it

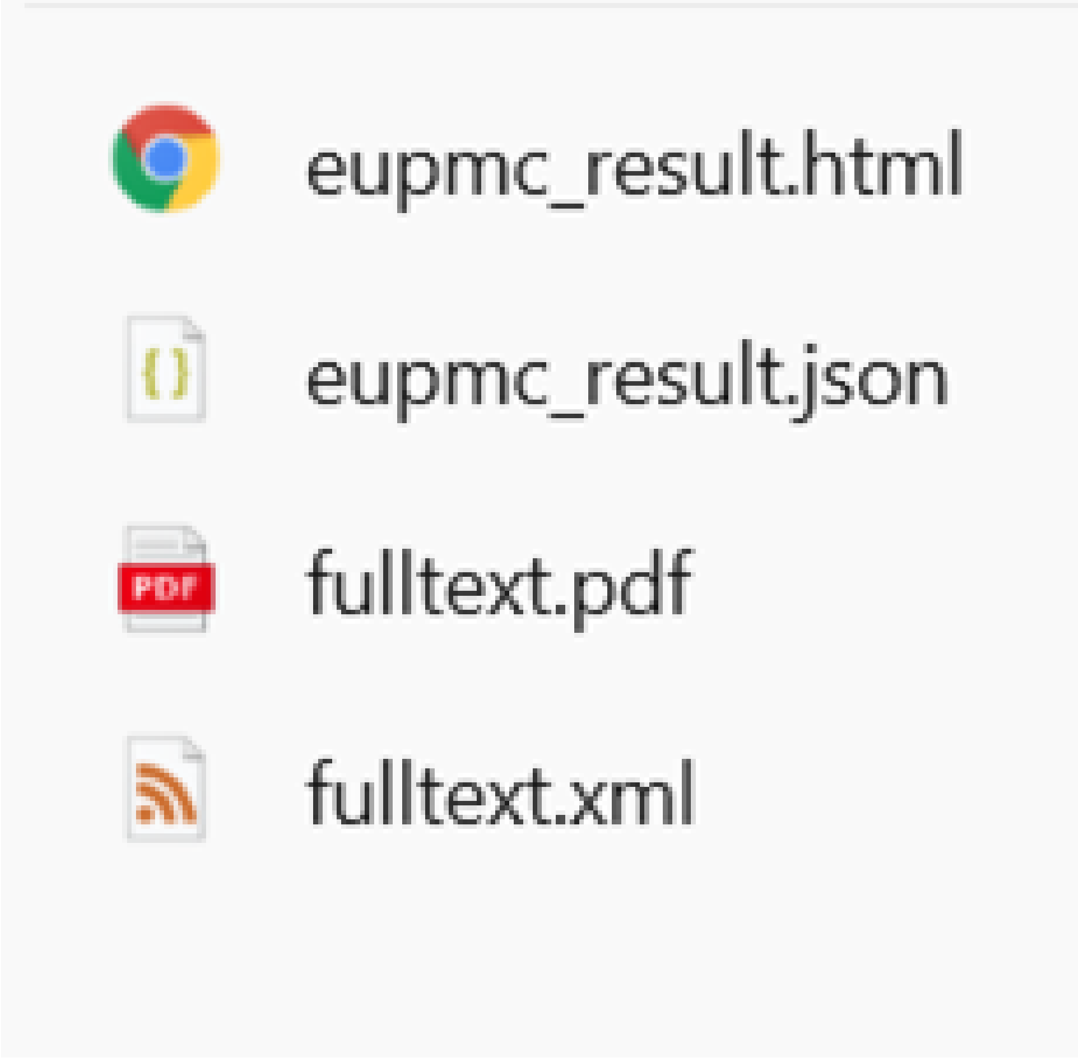
LOCAL CORPUS (CPROJECT)

Here we have a local corpus for the invasive species Lantana Camara for analysis



eupmc_result.html
eupmc_results.json
PMC8110560
PMC8112658
PMC8161263
PMC8190976
PMC8310452

Output Directory



eupmc_result.html
eupmc_result.json
fulltext.pdf
fulltext.xml

Individual Paper Folder

PYGETPAPERS DEMONSTRATION + LINKS

https://colab.research.google.com/drive/11iAbKQkFj0R6F9ZAX7ta1dSusJzXpiTX?usp=sharing

THANK YOU

AYUSH GARG-PYGETPAPERS