

# Hydra Scraper

Comprehensive scraper for **paginated APIs**,  
RDF, XML, file dumps, and Beacon files



[github.com/digicademy/hydra-scraper](https://github.com/digicademy/hydra-scraper)

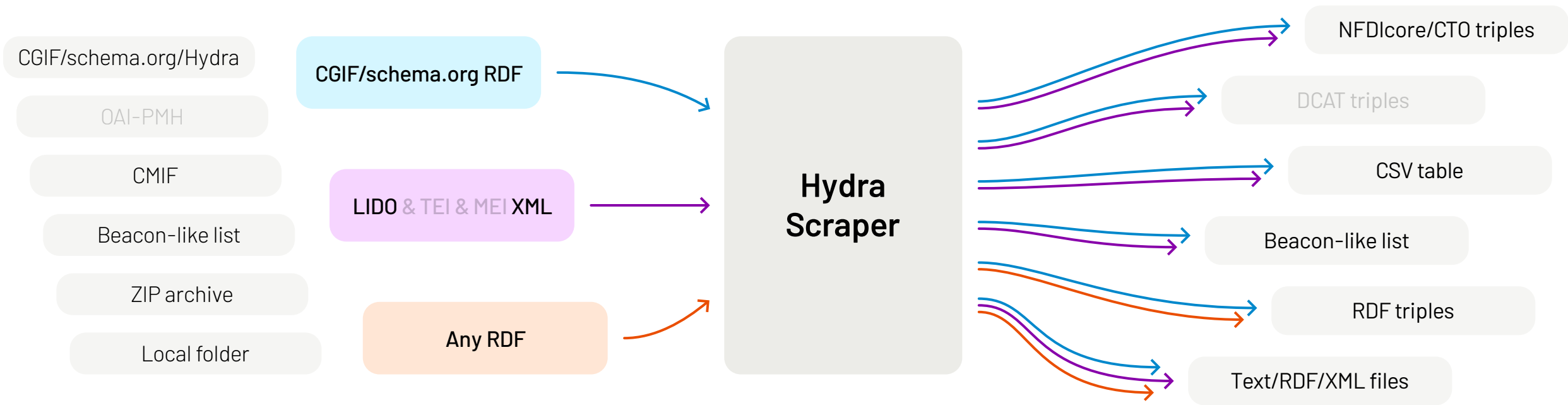
Originally developed for the **Corpus Vitrearum Germany**

One of the harvesters for the **Culture Knowledge Graph**

Command-line tool written in **Python**

Handy to check or harvest or convert **RDF or XML** data

## INPUT AND OUTPUT



**-l**

**-location** 1

*remote URL or local path*

**-f**

**-feed** 1

**-e**

**-elements** 1

**-o**

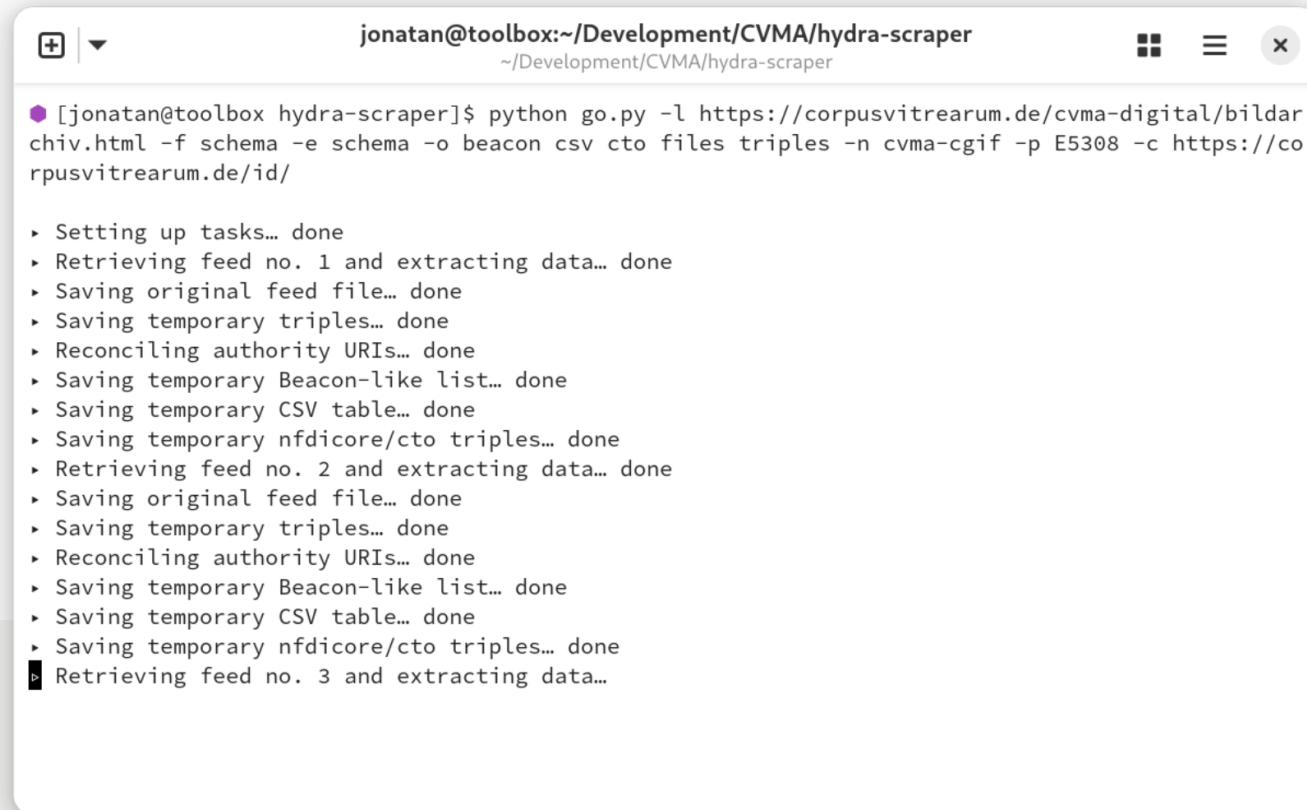
**-output** n

beacon  
cmif  
folder  
oaipmh  
schema  
schema-list

lido  
mei  
schema  
tei  
none

beacon  
csv  
cto  
files  
triples

## SAMPLE COMMAND



A terminal window titled 'jonatan@toolbox:~/Development/CVMA/hydra-scraper' with a subtitle '~/.Development/CVMA/hydra-scraper'. The window shows the execution of a Python script 'go.py' with various command-line arguments. The output is a list of progress messages, each preceded by a right-pointing triangle. The messages indicate the completion of tasks such as setting up tasks, retrieving and extracting data from three feeds, saving original feed files, saving temporary triples, reconciling authority URIs, saving temporary Beacon-like lists, saving temporary CSV tables, and saving temporary nfdicore/cto triples. The terminal is currently on the line 'Retrieving feed no. 3 and extracting data...'.

```
jonatan@toolbox:~/Development/CVMA/hydra-scraper  
~/.Development/CVMA/hydra-scraper  
[jonatan@toolbox hydra-scraper]$ python go.py -l https://corpusvitrearum.de/cvma-digital/bildarchiv.html -f schema -e schema -o beacon csv cto files triples -n cvma-cgif -p E5308 -c https://corpusvitrearum.de/id/  
▸ Setting up tasks... done  
▸ Retrieving feed no. 1 and extracting data... done  
▸ Saving original feed file... done  
▸ Saving temporary triples... done  
▸ Reconciling authority URIs... done  
▸ Saving temporary Beacon-like list... done  
▸ Saving temporary CSV table... done  
▸ Saving temporary nfdicore/cto triples... done  
▸ Retrieving feed no. 2 and extracting data... done  
▸ Saving original feed file... done  
▸ Saving temporary triples... done  
▸ Reconciling authority URIs... done  
▸ Saving temporary Beacon-like list... done  
▸ Saving temporary CSV table... done  
▸ Saving temporary nfdicore/cto triples... done  
▸ Retrieving feed no. 3 and extracting data...
```

python go.py

-l <https://corpusvitrearum.de/cvma-digital/bildarchiv.html>

-f schema -e schema -o beacon csv cto files triples

-n cvma-cgif -p E5308 -c <https://corpusvitrearum.de/id/>