

# Hydra Scraper

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



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

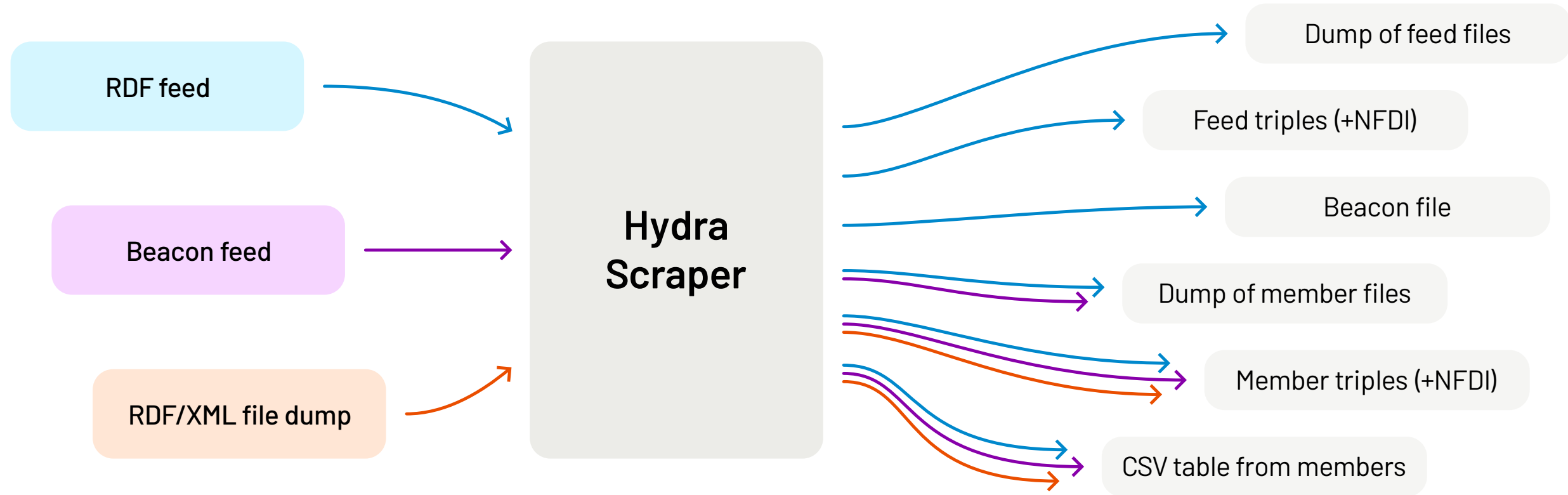
Originally developed for the **Corpus Vitrearum** Germany

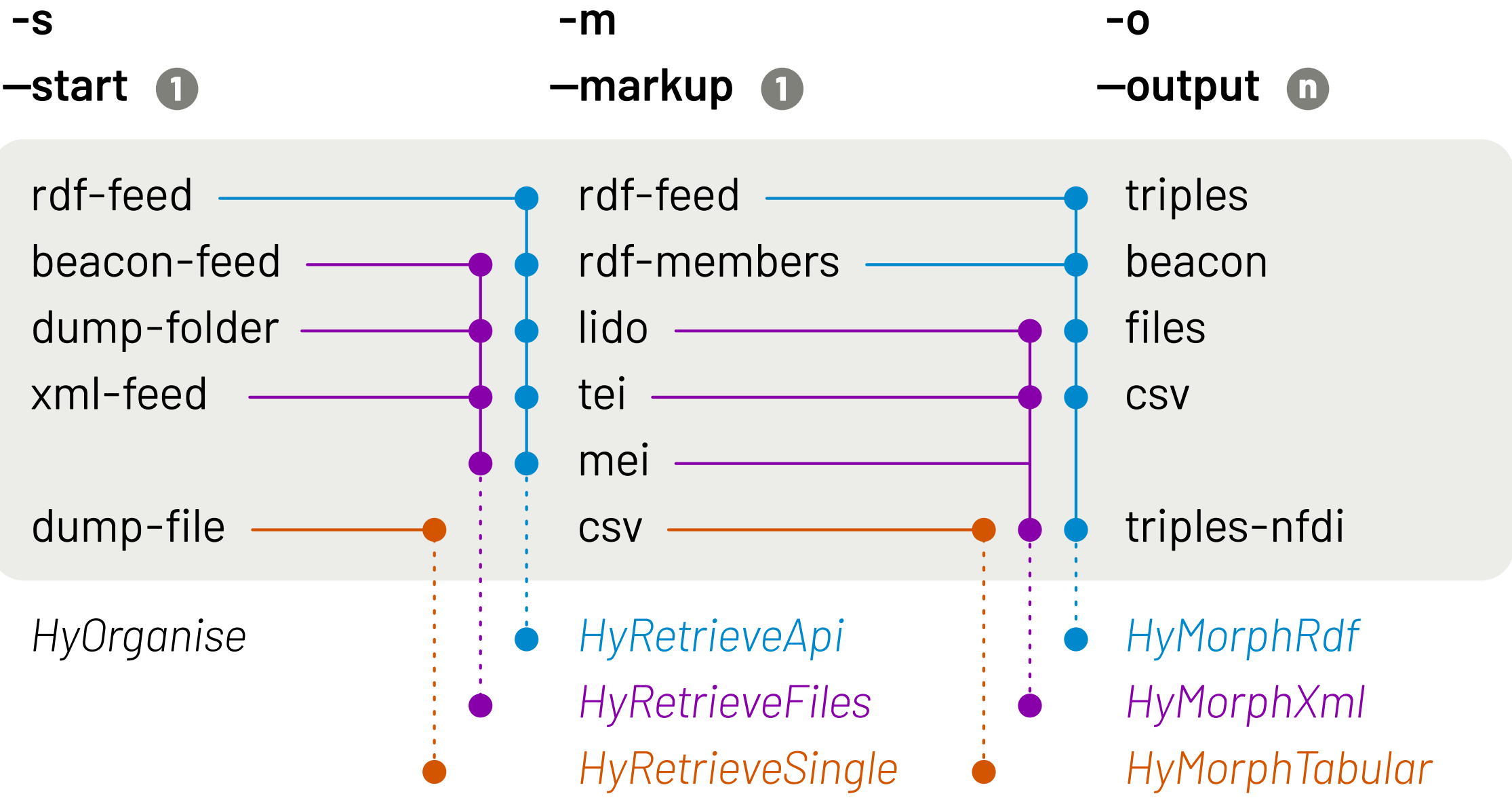
One of the harvesters for the **Culture Knowledge Graph**

Command-line tool written in **Python**

Currently handy to harvest or check **RDF** or **LIDO** data

## INPUT AND OUTPUT





## OLD SAMPLE COMMAND

```
jonatan@kuifje:~/Digitale Akademie/CVMA/hydra-scraper
~/Digitale Akademie/CVMA/hydra-scraper

[jonatan@kuifje hydra-scraper]$ python go.py -download 'lists,list_triples,list_cgif,beacon,r
esources,resource_triples,resource_cgif' -source_url 'https://corpusvitrearum.de/id/about.cgi
f' -target_folder 'cvma-cgif' -resource_url_filter 'https://corpusvitrearum.de/id/F' -resourc
e_url_add '/about.cgif' -clean_resource_names 'https://corpusvitrearum.de/id/,/about.cgif'

- Retrieving API lists... done!
- Saving beacon file... done!
- Saving list of API triples... done!
- Saving list of CGIF-filtered API triples... done!
- Retrieving individual resources... done!
- Saving list of resource triples... done!
- Saving list of CGIF-filtered resource triples... done!

Done! All lists saved to download folder. All resources listed in a beacon file. All API trip
les listed in a Turtle file. All CGIF-filtered API triples listed in a Turtle file. All resou
rces saved to download folder. All resource triples listed in a Turtle file. All CGIF-filtered
 resource triples listed in a Turtle file.

[jonatan@kuifje hydra-scraper]$
```

python go.py

-download 'lists,list\_triples,list\_cgif,beacon,resources,resource\_triples,resource\_cgif'  
-source\_url 'https://corpusvitrearum.de/id/about.cgif' -target\_folder 'cvma-cgif'  
-resource\_url\_filter 'https://corpusvitrearum.de/id/F' -resource\_url\_add '/about.cgif'  
-clean\_resource\_names 'https://corpusvitrearum.de/id/,/about.cgif'