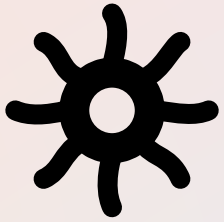


Hydra Scraper

JONATAN JALLE STELLER

ACADEMY OF SCIENCES AND LITERATURE MAINZ

MAY 2024, CC BY 4.0

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

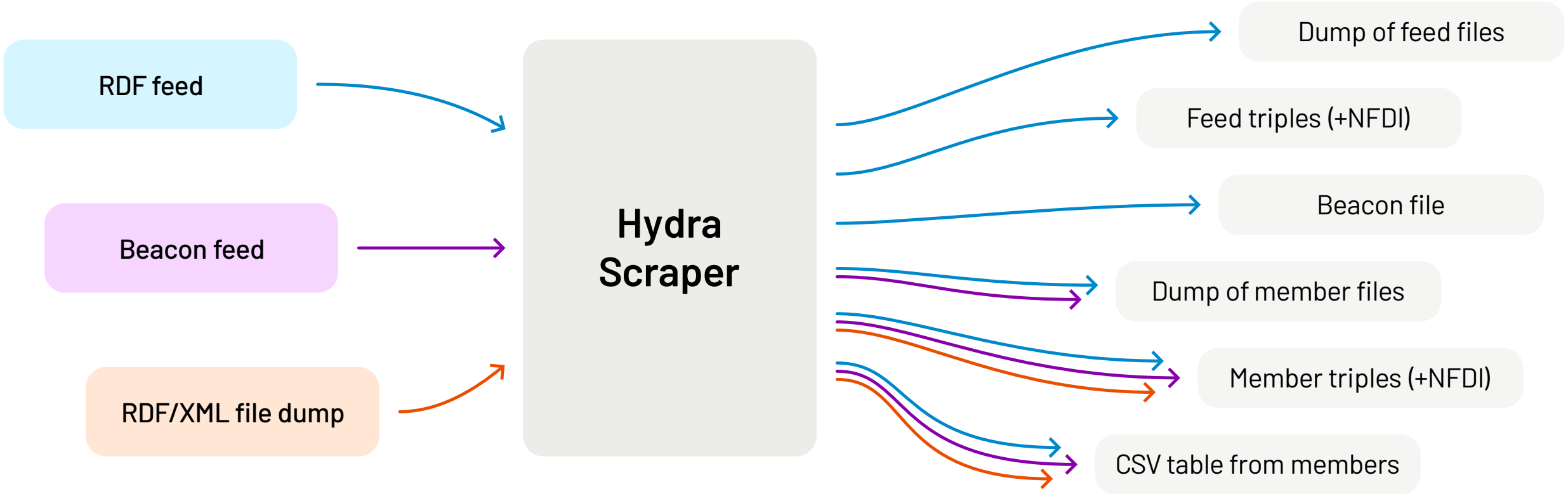
github.com/digicademy/hydra-scraper

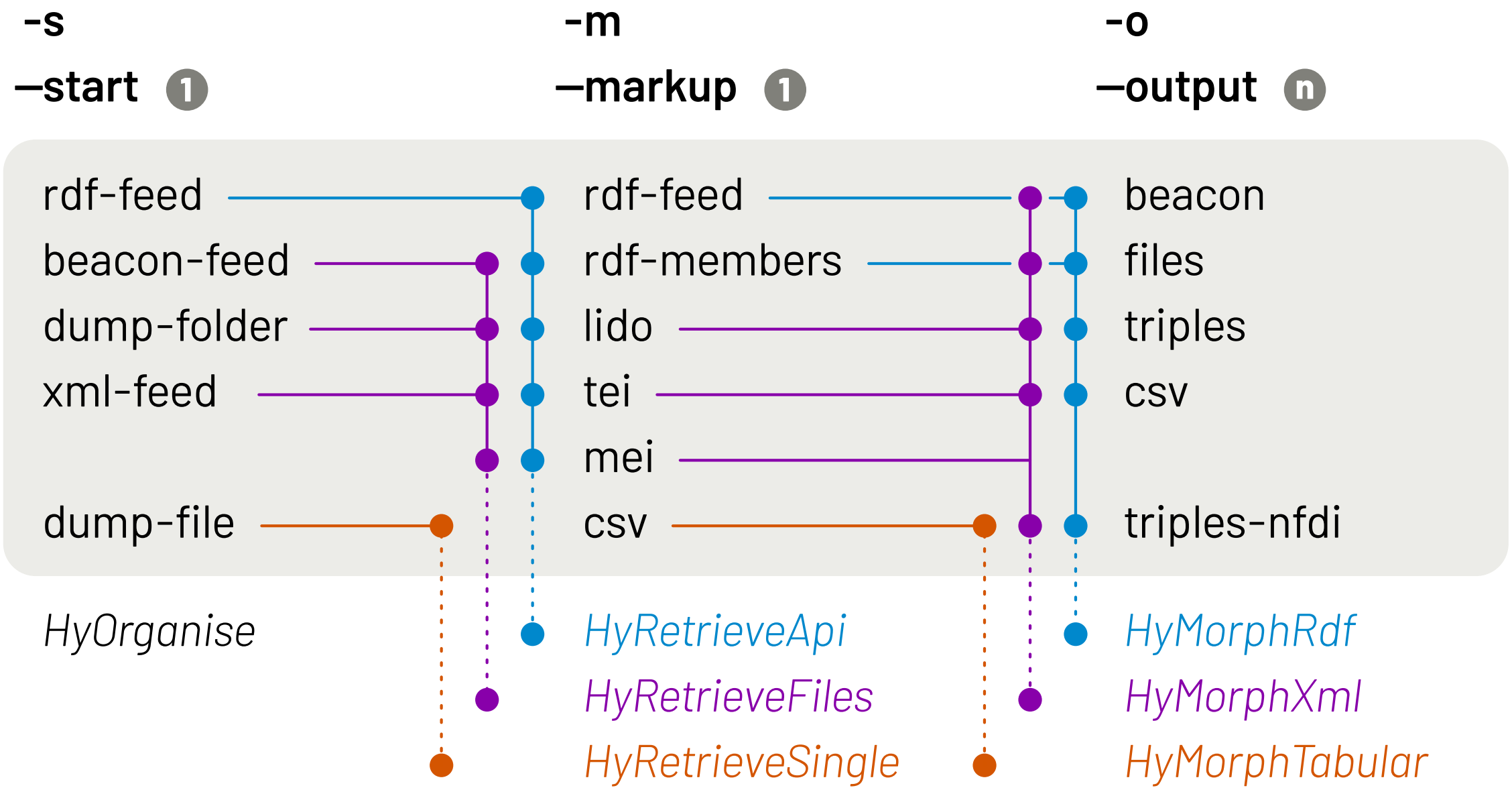
Originally developed for the **Corpus Vitrearum** Germany

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

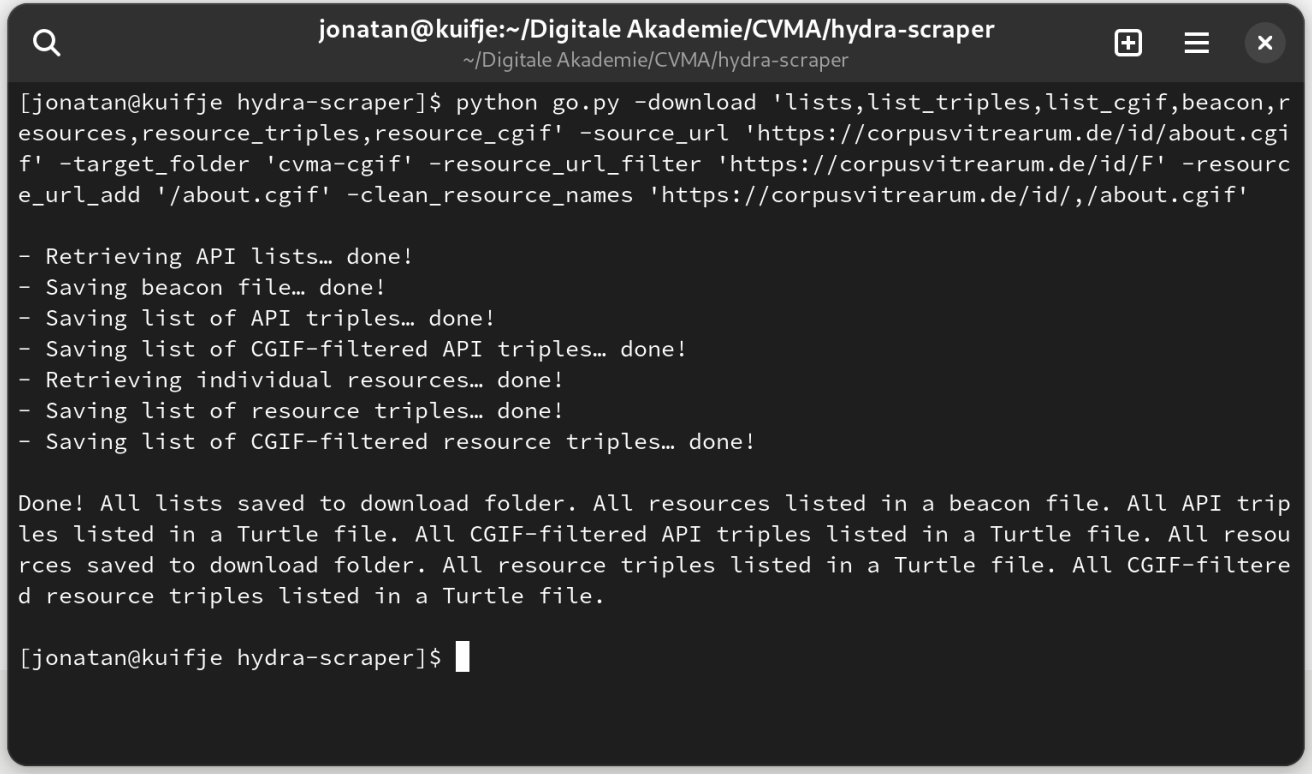
Command-line tool written in **Python**

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


RDF feed
Beacon feed
RDF/XML file dump
Hydra Scraper
Dump of feed files
Feed triples (+NFDI)
Beacon file
Dump of member files
Member triples (+NFDI)
CSV table from members


-s
–start 1
-m
–markup 1
-o
–output n
rdf-feed
beacon-feed
dump-folder
xml-feed
dump-file
rdf-feed
rdf-members
lido
tei
mei
csv
beacon
files
triples
csv
triples-nfdi
HyOrganise
HyRetrieveApi
HyRetrieveFiles
HyRetrieveSingle
HyMorphRdf
HyMorphXml
HyMorphTabular

OLD SAMPLE COMMAND



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'