Full-Text Resource Processing Training Workshop

Connecting text resource from Europeana Newspapers with CLARIN NLP tools using Jupyter notebooks

CLARIN ERIC & Europeana

15 June 2022

Welcome!

Full-Text Resource Processing Training Workshop

Organisers:

Michał Gawor, Twan Goosen (CLARIN ERIC)
Alba Irollo (Europeana Research)

Overview

14:00 - 14:15	Introductions
14:15 - 14:50	Tutorial (demo)
14:50 - 15:00	Break
15:00 - 15:45	Interactive tutorial (exercises)
15:45 - 16:00	Wrap-up

Introduction to Europeana

Alba Irollo

CLARIN

Common Language Resources and Technology infrastructure. A European Research Infrastructure Consortium (ERIC)

Provides easy and sustainable **access** to **language data** and **tools**

22 **national consortia**, 70 **connected centres**

Europeana partner (in Europeana DSI)

https://www.clarin.eu

NLP tools

CLARIN centres provide a wide range of natural language processing (NLP) tools

Tasks on text resources, e.g.:

- Parsing, Tagging, …
- Topic modelling, Named entity recognition, Sentiment analysis, …
- Translation, Text to Speech, Spelling correction…

https://www.clarin.eu/content/tools

Resources (language data)

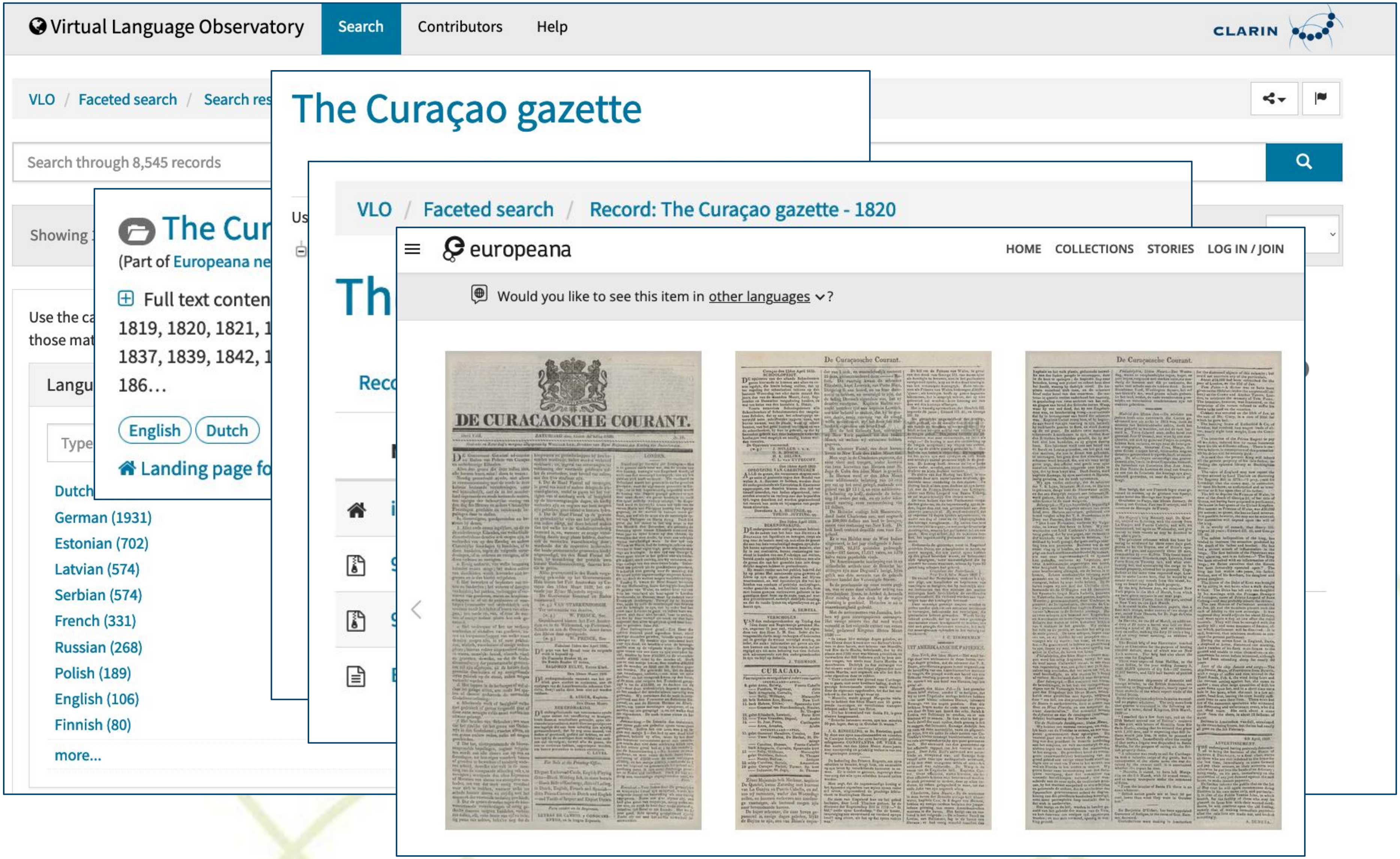
Many resource types can be distinguished:

- Textual, Audiovisual
- Corpora, Lexica, dictionaries, language models, ..

https://www.clarin.eu/content/data

CLARIN makes resources within and outside its consortium discoverable (vlo.clarin.eu).

Europeana newspaper full text resources have been aggregated at metadata level - easy to find and access via VLO.



Virtual Language Observatory | Search | Contributors | Help | CLARIN

VLO / Faceted search / Search res

Search through 8,545 records

Showing

Use the ca
those mat

Langu

Type

Dutch
German (1931)
Estonian (702)
Latvian (574)
Serbian (574)
French (331)
Russian (268)
Polish (189)
English (106)
Finnish (80)
more...

The Cur
(Part of Europeana ne

⊞ Full text conten
1819, 1820, 1821, 1
1837, 1839, 1842, 1
186…

English | Dutch

Landing page fo

The Curaçao gazette

Us

VLO / Faceted search / Record: The Curaçao gazette - 1820

Th

Reco

europeana | HOME | COLLECTIONS | STORIES | LOG IN / JOIN

Would you like to see this item in other languages ?

DE CURACAOSCHE COURANT.

De Curaçaosche Courant.

De Curaçaosche Courant.

https://vlo.clarin.eu

Introduction to Jupyter and NLP demo

Michał Gawor

Break until 15:00

You will be assigned to break-out rooms

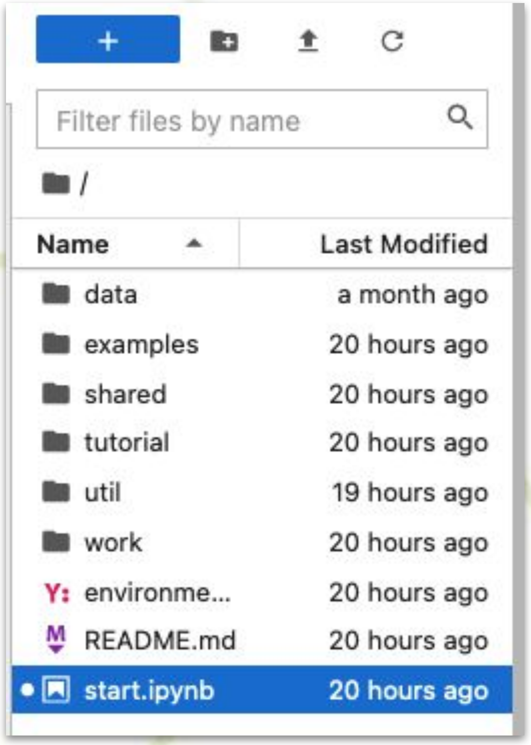
Please follow the log-in instructions
to access https://jupyter.clarin-dev.eu

Exercises

1. Go to the URL: jupyter.clarin-dev.eu

2. Log in with your personal credentials

3. Double-click on the ‘start.ipynb’ file on the left hand side and click the link to **Exercise 1** to get started



+

Filter files by name

/

Name	Last Modified
data	a month ago
examples	20 hours ago
shared	20 hours ago
tutorial	20 hours ago
util	19 hours ago
work	20 hours ago
environme...	20 hours ago
README.md	20 hours ago
start.ipynb	20 hours ago

15:40 Wrap-up in breakout

Wrap-up: summary and evaluation of the exercises

- Exercise 1
  - Notebook navigation and cell execution
- Exercise 2
  - Extracting information from metadata and resources files
- Exercise 3
  - Filtering data sets
  - Running NLP pipelines

Wrap-up: how to proceed?

- Environment at jupyter.clarin-dev.eu will remain available for one month
- Working with these and other notebooks after that:
  - SSHOC marketplace (sshopencloud.eu)
    -> search for *Jupyter notebooks Europeana newspapers*
  - One click to run in binder (but data will be lost)
  - Use a remote or local Jupyter environment
    - Local: anaconda.org (download + install)
    - Ask your institution!

Wrap-up: further exploration

- Using different data sets and/or tools
    - Europeana Newspaper full text currently available from 9 countries
    - CLARIN centres offer many different tools
        - Usage will not be the same for all tools
            - Calling from a notebook requires remote access
            - Get informed about means of access, e.g. web service or python library

- **See 'start.ipynb' notebook for links**

- **Experiment and ask for support in online communities**

Questions?

Thank you for your attention!

SSH Open Marketplace

Social Sciences and Humanities Open Marketplace

marketplace.sshopencloud.eu

CLARIN

Common Language Resources and
Technology Infrastructure

www.clarin.eu/notebooks

europeana

pro.europeana.eu/page/research

notebooks@clarin.eu