Teaching with CLARIN 2022
Introduction

Francesca Frontini, BoD
Iulianna van der Lek, Training Officer

CLARIN

11 October 2022
CLARIN Annual Conference

Programme

Introduction and Overview of Training Activities by Francesca Frontini and Iulianna van der Lek

Processing Europeana Text Collections with Jupyter Notebooks by Twan Goosen and Michal Gawor

Teaching with CLARIN Submissions 2022

- What's on the Agenda? Topic Modelling Parliamentary Debates before and during the COVID-19 Pandemic by Ajda Pretnar Žagar, Kristina Pahor de Maiti, Darja Fišer, Ljubljana, Slovenia
- Lithuanian Collocations: Usage, Teaching, Learning, and Translation by Jurgita Vaičenonienė, Vytautas Magnus University, Lithuania
- Natural Language Processing Methods by Rachele Sprugnoli, Dipartimento di Discipline Umanistiche, Sociali e delle Imprese Culturali – Università degli Studi di Parma, Italy

Discussion & Certificates

Teaching & Training | CLARIN KI

Knowledge Centres

Digital Humanities Course Registry

Tour de CLARIN

Teaching

Annual Conference

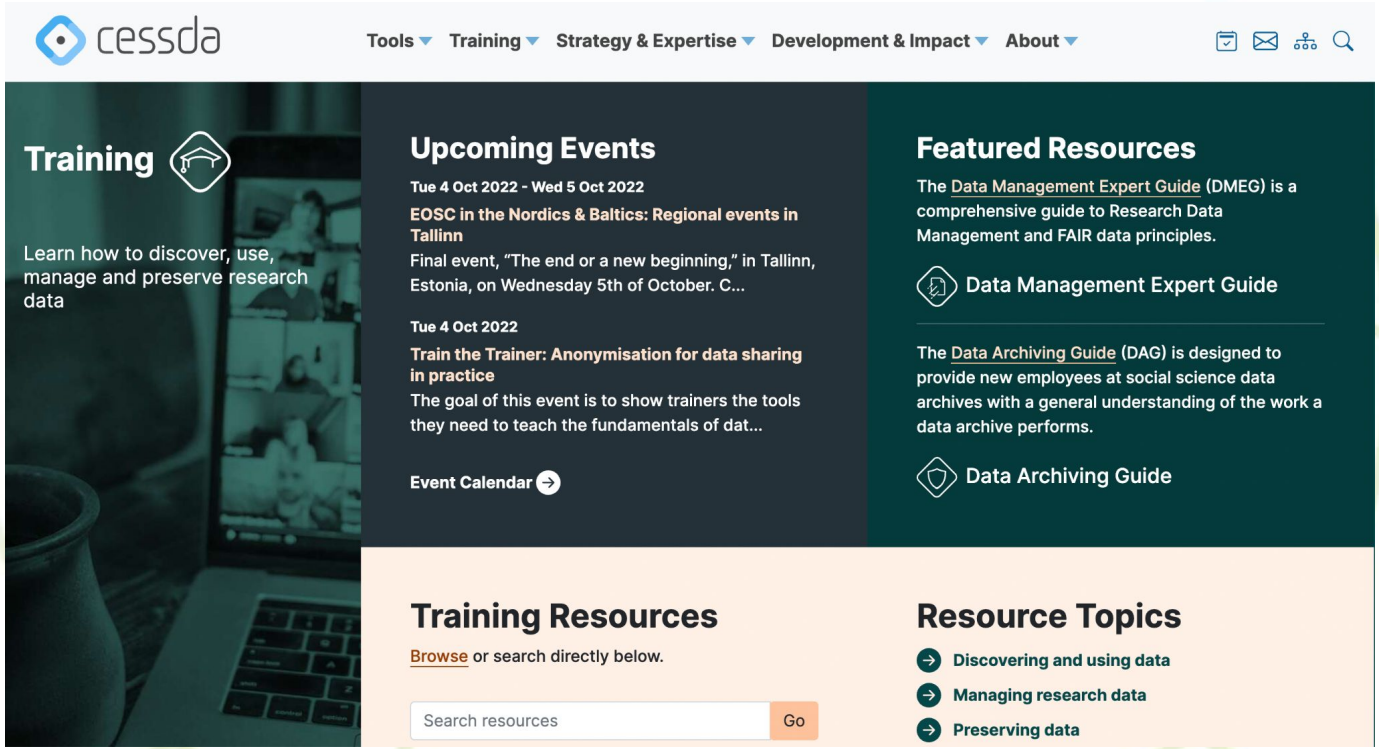
Funding

Video Channels

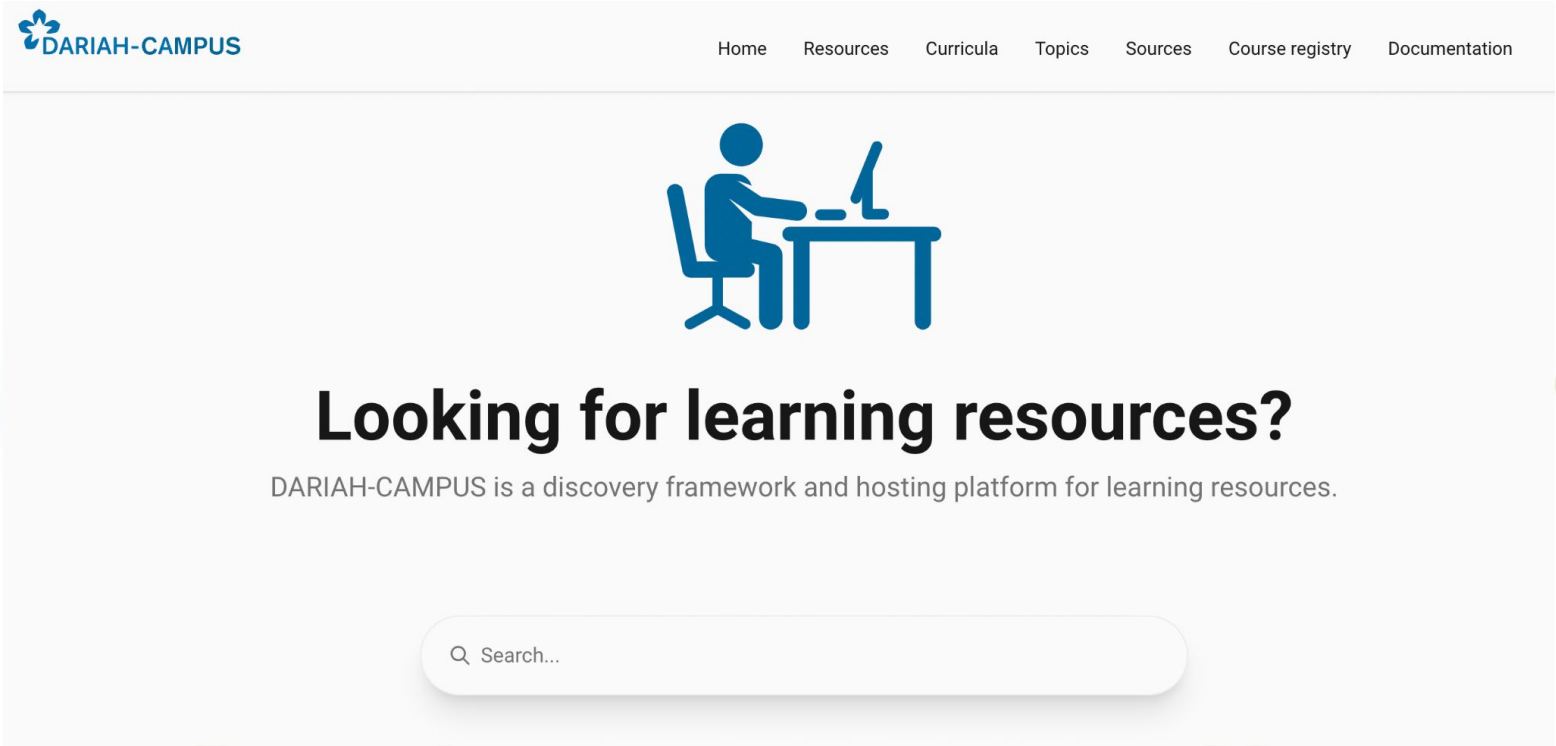
Best-Practice Papers

In alignment with other infrastructures

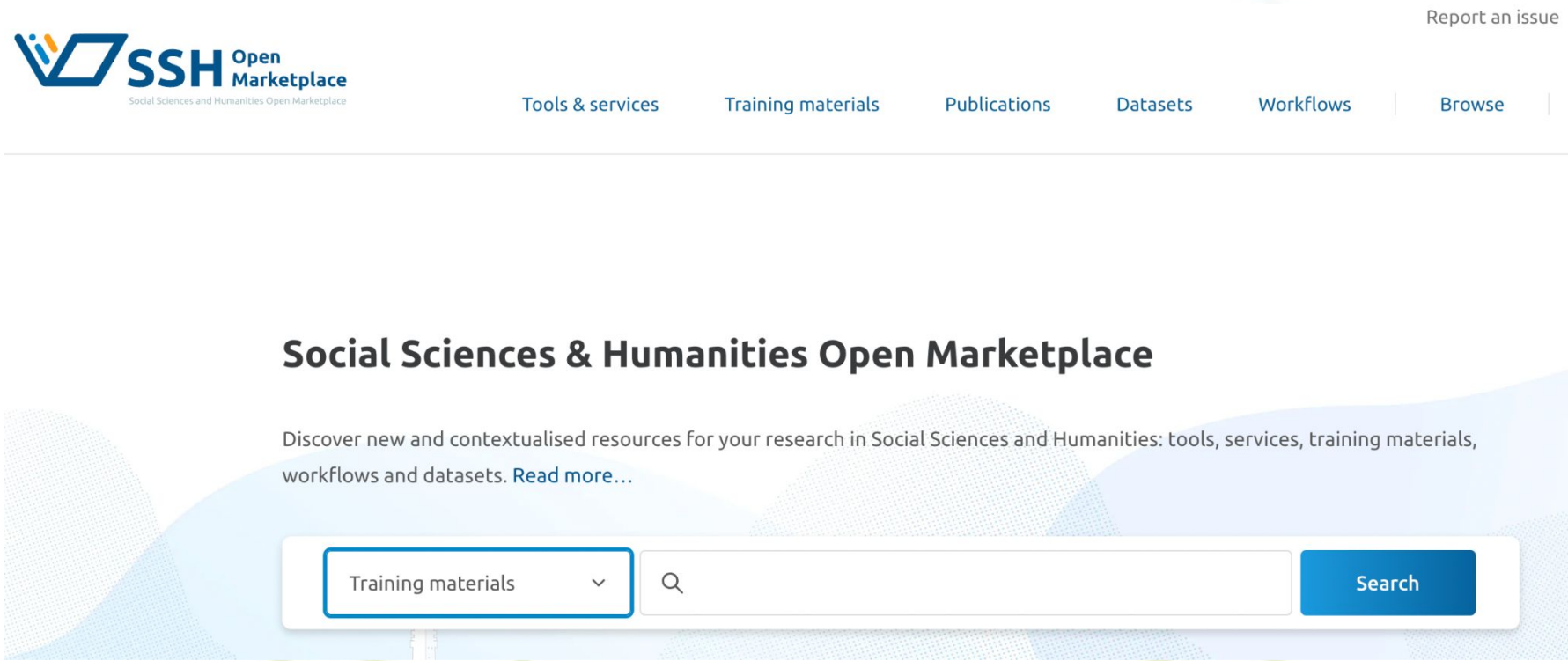
In alignment with other infrastructures



In alignment with other infrastructures



and with the SSHOC vision



2021: Teaching with CLARIN initiatives

Training Materials

Applied Language Technology
Author: Tuomo Hiippala
Faculty of Arts, University of Helsinki, Finland
Keywords: *language technology, digital humanities, tutorial, beginner, spaCy, Stanza, Universal Dependencies, introduction*

Archilochus of Paros: Elegiac Fragments - XML Archive
Author: Anika Nicolosi and Beatrice Nava
University of Parma, Italy
Keywords: *Ancient Greek, Fragmentary poetry, Textual criticism, Text annotation, Data science*

Computational Morphology with HFST
Author: Erik Axelson
Faculty of Arts, University of Helsinki, Finland
Keywords: *morphology, weighted finite-state networks, two-level rules, xfst, lexc, twolc*

GATE, an Open-Source Toolkit for Natural Language Processing
Author: Diana Maynard
Faculty of Engineering, University of Sheffield
Keywords: *Natural Language Processing; Machine Learning; GATE; social media analysis; disinformation; online abuse detection; Python; Deep Learning; information extraction; digital humanities; corpus linguistics; annotation*

Sharing teaching materials as digital objects

Description of the Training Materials

(Sub)discipline, topic, language(s)	**Topics**: Language technology, digital humanities **Language**: English
Keywords	language technology, digital humanities, tutorial, beginner, spaCy, Stanza, Universal Dependencies, introduction
Project URL	Applied Language Technology (mooc.fi) YouTube channel
CLARIN resources	The course materials build on various resources distributed through CLARIN, such as Universal Dependencies corpora. The materials refer to the CLARIN website for further study, highlighting the digital humanities course registry.
Structure and duration	The learning materials constitute a 10 ECTS module, consisting of two 5 credit courses. The materials are divided into two parts, in which each section corresponds to one week of studying.

Open issues:

- metadata
- visibility
- citability
- persistent identifiers
- versioning
- licensing

- FAIRness

CLARIN and FAIR Training Materials

CoP Task Force - OpenAire

Joint publication to be submitted in

[Scientific Data]

How to make your training materials FAIR

Abstract 2

Introduction 4
- The evolving FAIR landscape 4
- Research communities and FAIR 4
- FAIR training materials 5
- Research communities and FAIR training materials 7

Methodology 8
- Selection of case studies 8
- Structure of case studies and analysis 9

Implementing FAIR standards for training resources: lessons learned from case studies 10

Analysis and consolidation 10

Discussion 11

Further reading and appendices 19
- CLARIN 19
  - Context, scope and objectives 19
  - Implementation 19
  - Challenges and Lessons Learnt 20
- ELIXIR (work in progress! Just bullets for now) 23
  - ELIXIR Training 23
  - ELIXIR and FAIR training 23
  - Implementation 24
  - Challenges and Lessons learnt 25
- EMBL Bio-IT 26
  - Bio-IT: introduction 26
  - Scope and objectives 26
  - Implementation 26
  - Lessons learnt 27
- OpenAIRE 28
- SSHOC - Training Discovery Toolkit 30
- CESSDA Training resources catalogue- work in progress 33
- FORRT (work in progress) 35
- Context, scope and objectives 35
- Implementation 36
- Challenges and Lessons Learnt 38
- EOSC Synergy (in progress) 39
  - Background 39

1

2022: Teaching CLARIN

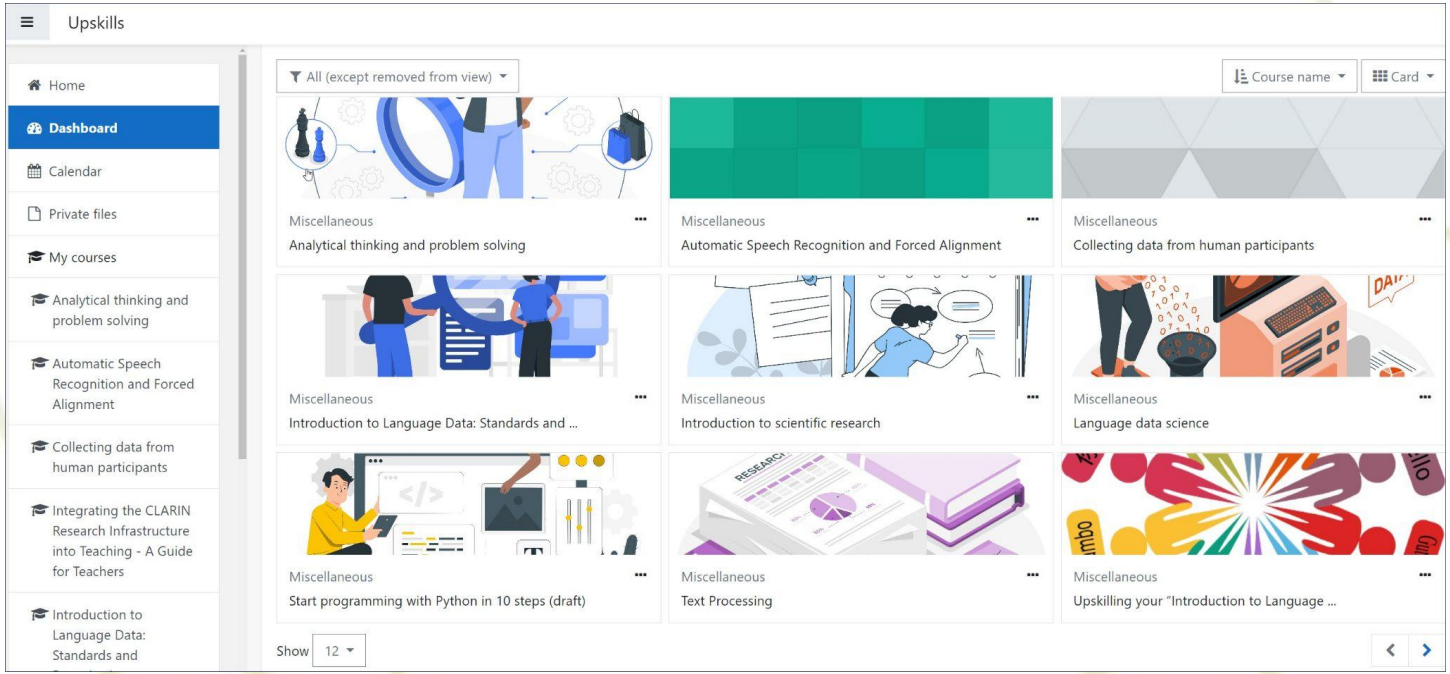
Developing new learning content to help teachers and lecturers to integrate CLARIN into their courses.

1) Focus on core CLARIN services:
   - VLO
   - Switchboard
   - Federated content search
   - Deposit and Citation

2) Automatic Speech Recognition & Force Alignment

What's next

Finalise the UPSKILLS learning content by February 2023 -> Testing, translation & piloting

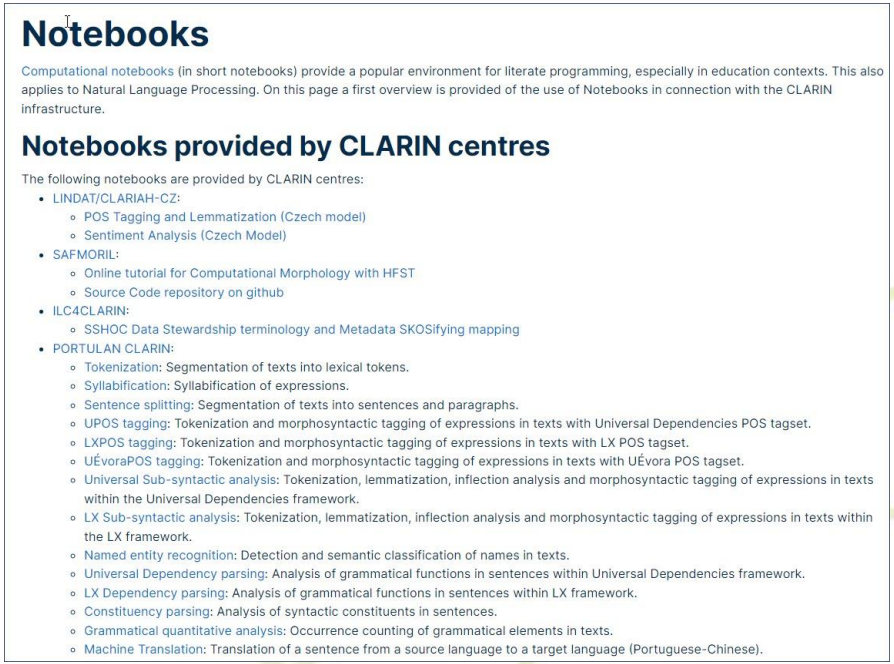


What's next

Student internships with FORLI from the University of Bologna

What's next

More text processing tutorials with Jupyter Notebooks



Notebooks

Computational notebooks (in short notebooks) provide a popular environment for literate programming, especially in education contexts. This also applies to Natural Language Processing. On this page a first overview is provided of the use of Notebooks in connection with the CLARIN infrastructure.

Notebooks provided by CLARIN centres

The following notebooks are provided by CLARIN centres:

- LINDAT/CLARIAH-CZ:
  - POS Tagging and Lemmatization (Czech model)
  - Sentiment Analysis (Czech Model)
- SAFMORIL:
  - Online tutorial for Computational Morphology with HFST
  - Source Code repository on github
- ILC4CLARIN:
  - SSHOC Data Stewardship terminology and Metadata SKOSifying mapping
- PORTULAN CLARIN:
  - Tokenization: Segmentation of texts into lexical tokens.
  - Syllabification: Syllabification of expressions.
  - Sentence splitting: Segmentation of texts into sentences and paragraphs.
  - UPOS tagging: Tokenization and morphosyntactic tagging of expressions in texts with Universal Dependencies POS tagset.
  - LXPOS tagging: Tokenization and morphosyntactic tagging of expressions in texts with LX POS tagset.
  - UÉvoraPOS tagging: Tokenization and morphosyntactic tagging of expressions in texts with UÉvora POS tagset.
  - Universal Sub-syntactic analysis: Tokenization, lemmatization, inflection analysis and morphosyntactic tagging of expressions in texts within the Universal Dependencies framework.
  - LX Sub-syntactic analysis: Tokenization, lemmatization, inflection analysis and morphosyntactic tagging of expressions in texts within the LX framework.
  - Named entity recognition: Detection and semantic classification of names in texts.
  - Universal Dependency parsing: Analysis of grammatical functions in sentences within Universal Dependencies framework.
  - LX Dependency parsing: Analysis of grammatical functions in sentences within LX framework.
  - Constituency parsing: Analysis of syntactic constituents in sentences.
  - Grammatical quantitative analysis: Occurrence counting of grammatical elements in texts.
  - Machine Translation: Translation of a sentence from a source language to a target language (Portuguese-Chinese).

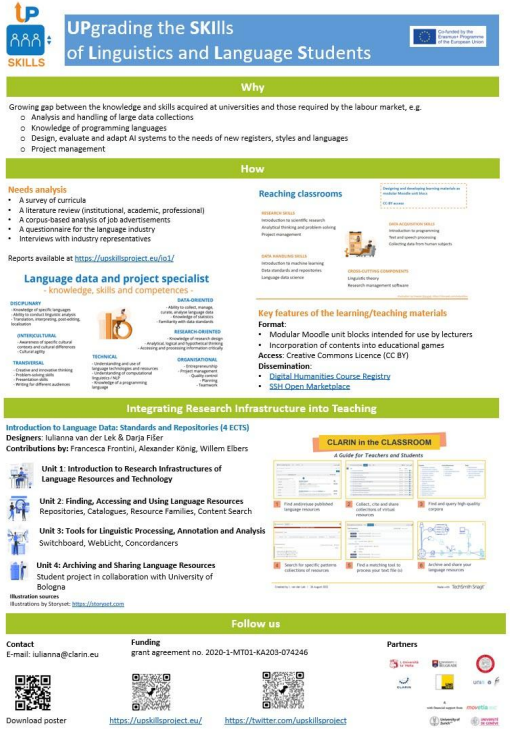
Next presentation

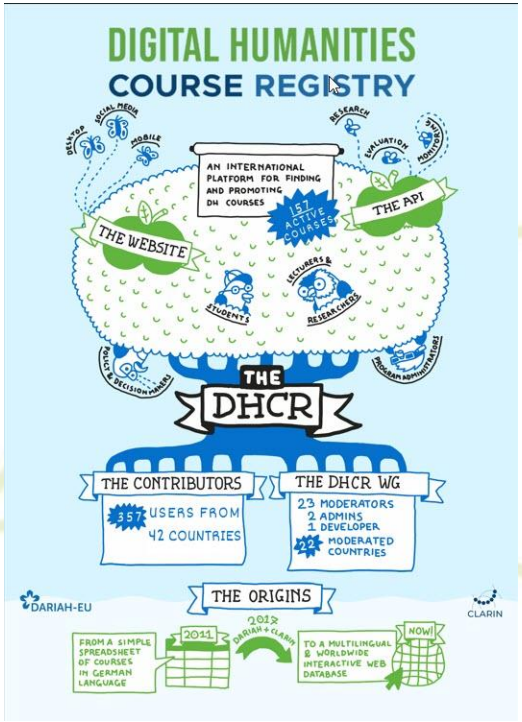
Processing Europeana Text Collections with Jupyter Notebooks by Twan Goosen and Michal Gawor

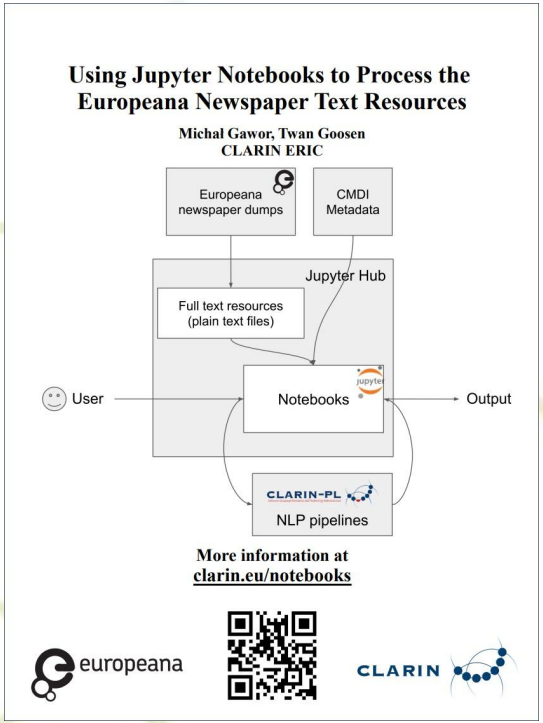
https://www.clarin.eu/notebooks

Would you like to know more?

Bazaar stall, today at 17:55 - 18:55







Get involved

Training and Education Task Force

training@clarin.eu

Subscribe to our trainers' network mailing list:

training@lists.clarin.eu