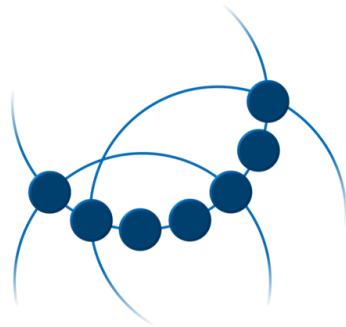


SSHOC Webinar

SSHOC'ing drama in the cloud

*Encoding theatrical text collections for discovery,
exploration, and visualisation;
the added value of SSHOC/CLARIN services*

CLARIN
Common Language Resources and
Technology Infrastructure



This project is funded from the EU Horizon 2020 Research and Innovation Programme (2014-2020) under Grant Agreement No. 823782



Housekeeping notes in the virtual space

- **The webinar is being recorded.** All participants will receive a link to the recording later today.
- **Slides are available:** See the chat box for the link.
- **Questions?** Put them in the chat box. The speakers will answer your questions at the end of the webinar.



SPEAKERS and the audience



Francesca Frontini
CLARIN ERIC Director & ILC CNR
SSHOC T3.1.
francesca.frontini@ilc.cnr.it



Maria Eskevich
CLARIN ERIC Central Office Coordinator
SSHOC T3.3. Task leader
maria@clarin.eu



Iulianna van der Lek
CLARIN ERIC Training and Education Officer
SSHOC T3.1.
iulianna@clarin.eu

What about you?

Please tell us who you are by answering a couple of questions shared with you via Zoom.

Structure of the tutorial

- 13.30 - 13.50: **Welcome and introduction**
 - CLARIN ERIC and What it Offers (Research Infrastructure, tools and services)
 - SSHOC (project details)
- 13.50 - 14.05: **Scenario of use and motivation**
 - a researcher with SSH research questions and limited knowledge of TEI
 - a librarian aware of SSHOC and CLARIN
- 14.05 - 14.25: **TEI in details**
- 14.25 - 14.35: Interaction with the audience

- 14.35 - 14.40 *Coffee-break (5 min)*

- 14.40 - 15.10: **SSHOC/CLARIN use case**
- 15.10 - 15.25: Questions & answers
- 15.25 - 15.30: Wrap-up and useful references

Introduction



CLARIN ERIC and What it Offers

www.clarin.eu



CLARIN in a nutshell

- **CLARIN** is the Common Language Resources and Technology Infrastructure
- **ESFRI ERIC** status since 2012, Landmark since 2016
- that provides easy and sustainable access for scholars in the **humanities and social sciences** and beyond
- to **digital language data** (in written, spoken, video or multimodal form)
- to **advanced tools** to discover, explore, exploit, annotate, analyse or combine them, wherever they are located
- to **education and training** webinars and workshops with focus on digital literacy of scholars, lecturers and students
- through a **single sign-on** environment
- that serves as an ecosystem for **knowledge exchange**
- and is ready for **integration in EOSC** (European Open Science Cloud; [link](#))

CLARIN Value Proposition: <https://www.clarin.eu/content/value-proposition> (link to [pdf](#))

CLARIN ERIC in members and centres

Consortium:

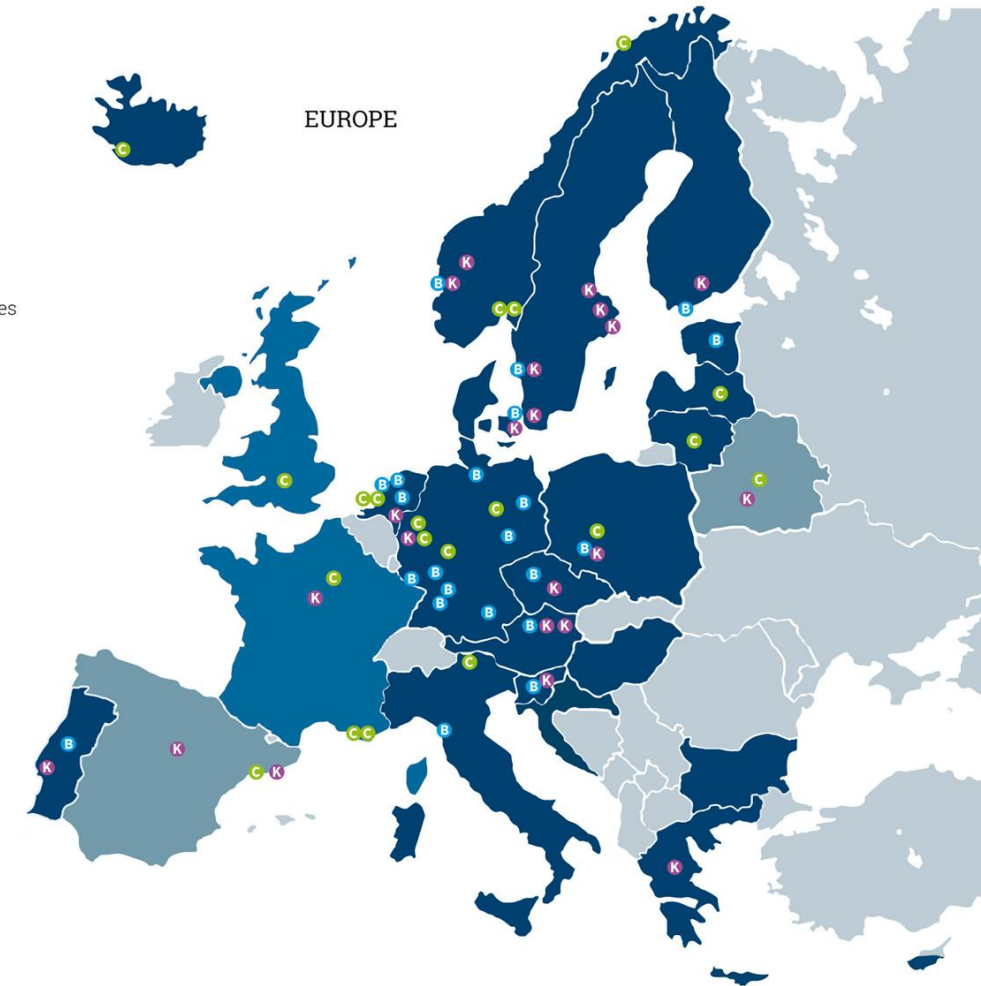
- 21 members: AT, BG, CY, CZ, DE, DK, EE, FI, GR, HR, HU, IS, IT, LT, LV, NL, NO, PL, PT, SE, SI
- 3 observers: FR, UK, ZA
- 1 linked party: CMU
- >60 centres
(incl. 25 CTS certified data centres)

Support for linguistic diversity

- Data covering more than 1500 languages
- Tools for many languages
- Language resources in all modalities



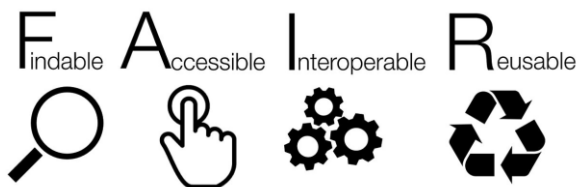
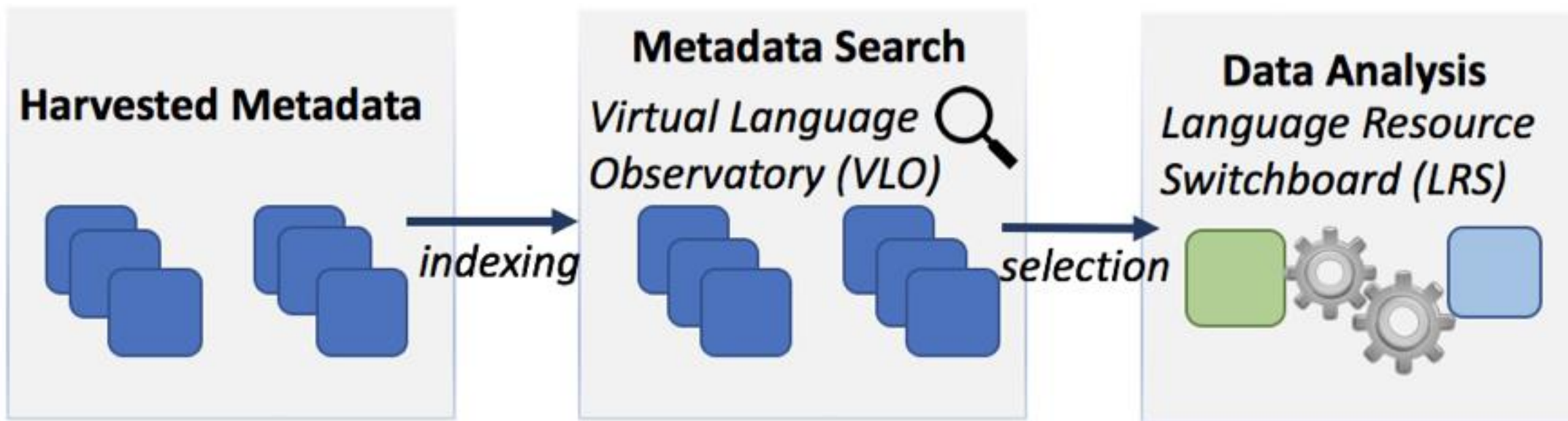
- ERIC members
- Observers
- Countries with participating centres
- B Centre Providing Data
- C Centre Providing Metadata
- K Knowledge Centre



CLARIN data types and communities of use in SSH and beyond

- Newspaper archives
- Parliamentary records
- Literary texts
- Historical letters
- Broadcast archives
- Oral History data
- Social Media data
- L-2 Learner Resources
- Survey data
- Patient recordings
- Excavation reports
- Digital humanities
- Linguistics and Philology
- Data Science /AI
- Translation and Lexicography
- Literary Studies
- History
- Political and Social Sciences
- Media Studies
- Culture, Folklore, Anthropology
- Speech therapy
- General Public

The CLARIN data architecture: *central processing of metadata*



clarin.eu/fair



vlo.clarin.eu




switchboard.clarin.eu






Search through 4,724 records

Showing 1 to 10 of 4,147 results within selection for Deutsches Textarchiv (1600–1900) x Results per page: 10 

Use the categories below to limit the search results to those matching the selected value(s).

Language Collection 

Deutsches Textarchiv (1600–1900) x


Modality Format Temporal Coverage Availability 

<< < 1 2 3 4 5 6 7 8 9 10 > >>

Neue Rheinische Zeitung

(Part of Deutsches Textarchiv (1600–1900))



 Historical German text source (16th–19th c.) digitized for the Deutsches Textarchiv (German Text Archive) collection according to the TEI-P5 guidelines.




German

[Landing page for this record](#)

Grundriß der Allgemeinen Volkswirtschaftslehre

(Part of Deutsches Textarchiv (1600–1900))



 Historical German text source (16th–19th c.) digitized for the Deutsches Textarchiv (German Text Archive) collection according to the TEI-P5 guidelines.



German

[Landing page for this record](#)

Reigen

(Part of Deutsches Textarchiv (1600–1900))





 Hilfe

 in den Titeldaten
 im Korpus
 in der Dokumentation


Text-Bild-Ansicht öffnen ...

Neue Rheinische Zeitung. Nr. 120. Köln, 19. Oktober 1848.

BIBLIOGRAPHISCHE ANGABEN

URN: urn:nbn:de:kobv:b4-31453-9
 Titel: Neue Rheinische Zeitung
 Untertitel: Organ der Demokratie
 weiterer Titel: Nr. 120, Donnerstag, 19. Oktober 1848
 Erscheinungsjahr: 1848
 Verlag/Drucker: Clouth
 Ort: Köln
 Auflage: 1. Auflage
 Bildnachweis: Russisches Staatsarchiv für sozio-politische Geschichte, RGASPI, Moskau, f. 1, op. 1, d. 268

ZUGEHÖRIGE WERKE

- Neue Rheinische Zeitung. Nr. 1. Köln, 1. Juni 1848.
- Neue Rheinische Zeitung. Nr. 2. Köln, 2. Juni 1848.
- Neue Rheinische Zeitung. Nr. 3. Köln, 3. Juni 1848.
- Neue Rheinische Zeitung. Nr. 4. Köln, 4. Juni 1848.
- Neue Rheinische Zeitung. Nr. 5. Köln, 5. Juni 1848.
- Neue Rheinische Zeitung. Nr. 6. Köln, 6. Juni 1848.
- Neue Rheinische Zeitung. Nr. 7. Köln, 7. Juni 1848.
- Neue Rheinische Zeitung. Nr. 7. Köln, 7. Juni 1848. Beilage.
- Neue Rheinische Zeitung. Nr. 8. Köln, 8. Juni 1848.
- Neue Rheinische Zeitung. Nr. 9. Köln, 9. Juni 1848.
- Neue Rheinische Zeitung. Nr. 9. Köln, 9. Juni 1848. Beilage.
- Neue Rheinische Zeitung. Nr. 10. Köln, 10. Juni 1848.
- Neue Rheinische Zeitung. Nr. 10. Köln, 10. Juni 1848. Beilage.
- Neue Rheinische Zeitung. Nr. 11. Köln, 11. Juni 1848.
- Neue Rheinische Zeitung. Nr. 11. Köln, 11. Juni 1848. Beilage.

Suche im Werk

 Hilfe

Ansichten für dieses Werk

- Text-Bild-Ansicht
- alle Faksimiles
- DTAQ (Qualitätssicherung)

Download

XML (TEI P5) · HTML · Text
 TCF (text annotation layer)
 TCF (tokenisiert, serialisiert, lemmatisiert, normalisiert)
 XML (TEI P5 inkl. att.linguistic)

Metadaten

TEI-Header · CMDI · Dublin Core

Statistiken

Scans: 4
 Zeichen: ca. 86 338
 Tokens: ca. 12 157
 Oberflächentypes: ca. 3 950

Wortwolken

- Lemmata
- Lemmata (nur Nomen)
- Types
- Types (nur Nomen)

https://www.deutschestextarchiv.de/book/show/nn_nrhz120_1848

National Library of Norway

Website	https://www.nb.no/sprakbanken/en/sprakbanken/
Consortium	CLARINO
Type(s)	C
Type status	Aiming for B
Description	Providing metadata services and access to language resources
CoreTrustSeal/DSA	<i>none</i>
PID status	Handle via EPIC (prefix: 21.11146)
Repository system	Custom
Strict versioning?	<input checked="" type="checkbox"/>

Organisational information

Organisation name	Språkbanken (Speech & Language Data Bank)
-------------------	---

Endangered Languages Archive

Website	https://www.elararchive.org/
Consortium	CLARIN-UK
Type(s)	C
CoreTrustSeal/DSA	<i>none</i>
Repository system	Preservica
Strict versioning?	<input checked="" type="checkbox"/>

Organisational information

Organisation name	Endangered Languages Archive
Institution	SOAS University of London
Working unit	SOAS Library
Shorthand	ELAR

Resources

nn_nrhz120_1848.TEI-P5.xml 106.14 KiB



Mediatype

text/plain

Language

German

Matching Tools

Group by task

Search for tool

▼ Constituency Parsing



Open 

WebLicht Const Parsing DE 

▼ Dependency Parsing



Open 

Spacy (hosted by D4Science) - DE



Open 

UDPipe



Open 

WebLicht Dep Parsing DE 

▼ Distant Reading



Open 

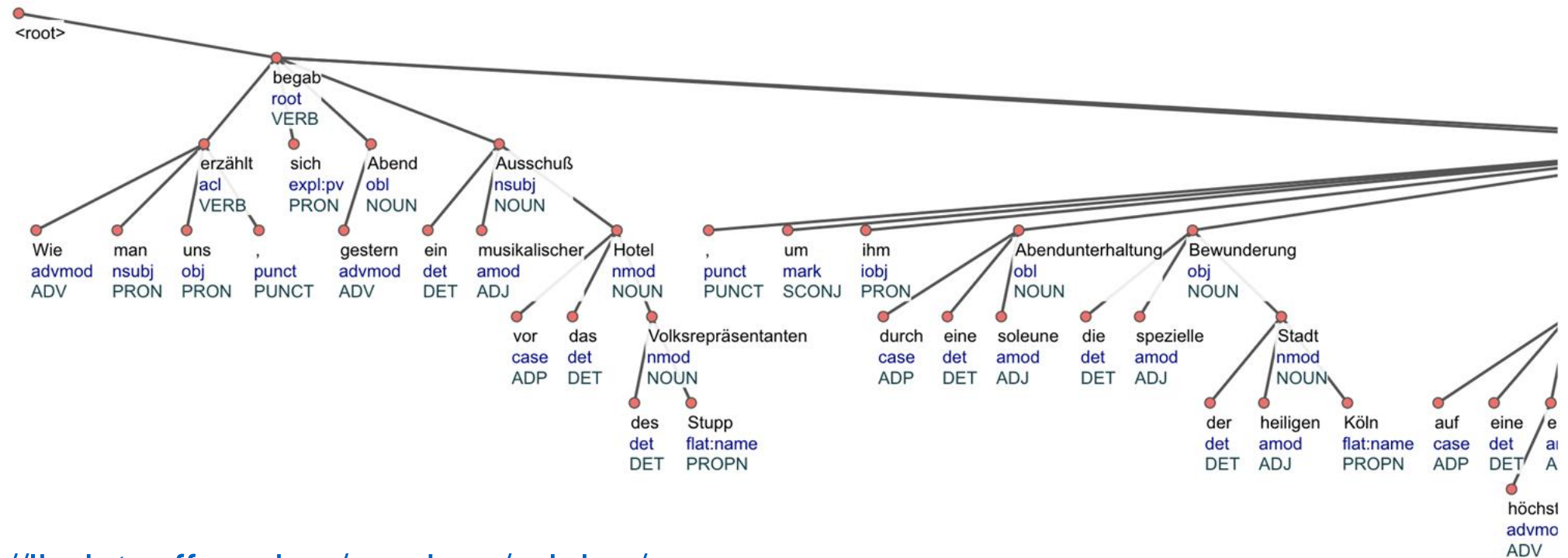
Voyant Tools

switchboard.clarin.eu

UDPipe

[About](#)
[Run](#)
[REST API Documentation](#)
[Save Tree as SVG](#)

Wie man uns erzählt , begab sich gestern Abend ein musikalischer Ausschuß vor das Hotel des Volksrepräsentanten Stupp , um ihm durch eine soleune Abendunterhaltung die spezielle Bewunderung der heiligen Stadt Köln auf eine höchst eklatante Weise zu erkennen zu geben .



<https://lindat.mff.cuni.cz/services/udpipe/>

CLARIN as Ecosystem for Knowledge Exchange



Knowledge centres

CLARIN Knowledge Centres (abbreviated K-centres) are a cornerstone of the CLARIN knowledge infrastructure.

K-centres are institutions that have agreed to share their knowledge and expertise on one or more aspects of the domain covered by the CLARIN infrastructure.

<https://www.clarin.eu/content/knowledge-infrastructure>

CLARIN as Ecosystem for Knowledge Exchange



Knowledge centres

- Individual languages (e.g. Danish, Czech, Portuguese), language families (e.g. South Slavic) or groups of languages (e.g. morphologically rich languages, the languages of Sweden)
- Written text and modalities other than written text (e.g. spoken language, sign language)
- Linguistic topics (e.g. language diversity, language learning, diachronic studies)
- Language processing topics (e.g. speech analysis, building treebanks, machine translation)
- Data types other than corpora (e.g. lexical data, word nets, terminology banks)
- Using or processing families of language data that will exist for most languages (e.g. newspapers, parliamentary records, oral history)
- Generic methods and issues (e.g. data management, ethics, IPR, OCR)

IMPACT-CKC

Areas of competence

Audiences served

Types of services

Is portal for language(s)

Other languages covered

Modalities covered

Linguistic topics

Language processing

Data types

Resource families

Generic topics

Other keywords

Tour de CLARIN

IMPACT centre of competence - CLARIN K-centre in digitisation

IMPACT-CKC (IMPACT centre of competence - CLARIN K-centre in digitisation), as knowledge centre offers expertise and resources to institutions and researchers looking for advice in digitisation and related fields. The IMPACT-CKC resources include a demonstrator platform for online testing tools, a collection of high quality images with associated ground truth, historical lexica for 10 languages as well as training materials and registries on tools, initiatives, datasets and competitions.

- researchers; - **librarians**; - archivists; - digital humanists; - computer scientists in topics related to digitisation

- Access to data; - Access to tools; - Training; - User assistance

-

- Spanish; - English; - Polish; - French; - Dutch; - German; - Slovene; - Czech; - Latin; - Bulgarian

- Audio-visual; - Text

- corpus linguistics; - diachronic language resources; - language learning

- basic language processing; - information extraction

- lexical data; - language models; - linked open data; - ontologies

- Historical corpora; - Lexica; - Literary corpora; - Newspaper corpora

- OCR; - digitisation; - visualisation; - evaluation of tools

-

[Introduction Interview](#)

CLARIN as Ecosystem for Knowledge Exchange



HOME • BROWSE LECTURES • PEOPLE • CONFERENCES • ACADEMIC ORGANISATIONS • EU SUPPORTED • ABOUT US SEARCH >>


Event: [Academic Organisations](#) » [CLARIN - Common Language Resources and Technology Infrastructure](#) » Doing text analytics for Digital Humanities and Social Sciences with CLARIN (LDK tutorial), Galway 2017

View order

- Overview
- Hot
- Popular
- Just published
- Recent
- Top Voted



LDK tutorial 2017 - Galway

CLARIN 

Watching Now

This page - Doing text analytics for Digital Humanities and Social Sciences with CLARIN (LDK tutorial), Galway 2017

SHOW CHAT >>

Doing text analytics for Digital Humanities and Social Sciences with CLARIN (LDK tutorial), Galway 2017

released under terms of: [Creative Commons Attribution \(CC-BY\)](#)

Text is a basic material, a primary data layer, in many areas of humanities and social sciences. If we want to move forward with the agenda that the fields of digital humanities and computational social sciences are projecting, it is vital to bring together the technical areas that deal with automated text processing, and scholars in the humanities and social sciences. Much progress has been made in the last two decades in text analytics, a field that draws on recent advances in computational linguistics, information retrieval and machine learning. By now we know what to expect from basic tools, such as named entity recognition. To foster new areas of research, it is necessary to not only understand what is out there in terms of proven technologies and infrastructures such as CLARIN, but also how the developers of text analytics can work with researchers in the humanities and social sciences to understand the challenges in each other's field better. What are the research questions of the researchers working on the texts? Can answering these questions be supported by computational models (in a non-reductionistic way)?

The LDK tutorial took place on 18 June 2017, as part of the preconference programme for LDK 2017, the conference on Language, Data and Knowledge that took place on 19-20 June 2017 in Galway, Ireland. The tutorial is co-organized by CLARIN and DARIAH-Ireland.

Categories

- Top » Humanities
- Top » Social Sciences
- Top » Computers » Digital Media

HOME • BROWSE LECTURES • PEOPLE • CONFERENCES • ACADEMIC ORGANISATIONS • EU SUPPORTED • ABOUT US SEARCH >>

Project: [Academic Organisations](#) » [CLARIN - Common Language Resources and Technology Infrastructure](#)

View order

- Overview
- Hot
- Popular
- Just published
- Recent
- Top Voted




ParlaCLARIN Workshop: Creating and Using Parliamentary Corpora, Miyazaki 2018

ParlaCLARIN Workshop 2018 - Miyazaki

CLARIN - COMMON LANGUAGE RESOURCES AND TECHNOLOGY INFRASTRUCTURE

RECENT EVENTS [MORE...](#)

CLARIN 

Watching Now

- This page - CLARIN - Common Language Resources and Technology Infrastructure
- Category: Computers
- International Center for Advanced Video Lectures.NET
- Course on Information Theory, Pattern
- PhD Thesis Defense: Dynamics of large

CLARIN stands for "Common Language Resources and Technology Infrastructure".

It is a research infrastructure that was initiated from the vision that all digital language resources and tools from all over Europe and beyond are accessible through a single sign-on online environment for the support of researchers in the humanities and social sciences.

In 2012 CLARIN ERIC was established and took up the mission to create and maintain an infrastructure to support the sharing, use and sustainability of language data and tools for research in the humanities and social sciences. Currently CLARIN provides easy and sustainable access to **digital language data** (in written, spoken, or multimodal form) for scholars in the **social sciences** and **humanities**, and beyond. CLARIN also offers **advanced tools** to discover, explore, exploit, annotate, analyse or combine such data sets, wherever they are located. This is enabled through a networked federation of centres: language data repositories, service centres and knowledge centres, with single sign-on access for all members of the academic community in all participating countries. Tools and data from different centres are interoperable, so that data collections can be combined and tools from different sources

CLARIN Annual Conference 2019 - Leipzig

8th CLARIN Annual Conference, Leipzig 2019

The CLARIN Annual Conference is the main annual event for those working on the construction and operation of CLARIN across ...

2nd Baltic Summer School of Digital Humanities

Essentials of Coding and Encoding

Baltic Summer School of Digital Humanities

Essentials of Coding and Encoding

CLARIN as Ecosystem for Knowledge Exchange

A banner with a dark background and blurred text, featuring the title 'CLARIN Resource Families' in white, bold, sans-serif font.

CLARIN Resource Families

Get to know the different types of resources in different languages CLARIN makes available for researchers from digital humanities, social sciences and human language technologies.

Corpora

- Computer-mediated communication corpora
- Corpora of academic texts
- Historical corpora
- L2 learner corpora
- Literary corpora
- Manually annotated corpora
- Multimodal corpora
- Newspaper corpora
- Parallel corpora
- Parliamentary corpora
- Reference corpora
- Spoken corpora

Lexical Resources

- Lexica
- Dictionaries
- Conceptual Resources
- Glossaries
- Wordlists

Tools

- Normalization
- Named entity recognition
- Part-of-speech tagging and lemmatization
- Tools for sentiment analysis

<https://www.clarin.eu/resource-families>

Historical corpora in the CLARIN infrastructure

Monolingual corpora

Corpus	Language	Description	Availability
Open Richly Annotated Cuneiform Corpus, Korp Version Size: 741,100 tokens Annotation: tokenised Licence: CC-BY-SA	Akkadian	This corpus contains cuneiform texts from Ancient history. The corpus is available through the concordancer Korp.	Concordancer
Greek Medieval Texts Size: 3.4 million words Licence: CC-BY	Ancient Greek	This corpus contains texts from the 4th to the 16th century. The corpus is available for download from the clarin:el repository.	Download
Sheffield Corpus of Chinese Annotation: no annotation Licence: CC-BY-NC-SA 3.0	Chinese	This corpus contains fictional and non-fictional texts from the Medieval and Modern Chinese periods. The corpus is available for download from the Oxford Text Archive.	Download

TABLE OF CONTENTS

[Introduction](#)

Historical corpora in the CLARIN infrastructure

[Monolingual corpora](#)

[Multilingual corpora](#)

[Other historical corpora](#)

[Additional materials](#)

[List of publications on historical corpora](#)

ParlaMint: Towards Comparable Parliamentary Corpora

Motivation| Mission and Goal| Expected Outcome| Tasks| Workplan| Results| Observers| Dissemination| Events| Participants| Financial Support| Contacts

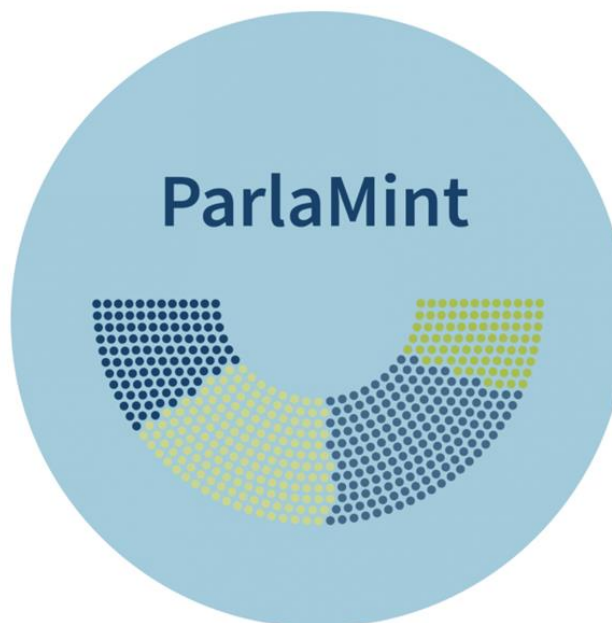
Motivation

National parliamentary data is a verified communication channel between the elected political representatives and society members in any democracy. It needs to be made accessible and comprehensive - especially in times of a global crisis. With the recent advances of artificial intelligence, analytics over unstructured parliamentary data for many languages is rapidly becoming a prerequisite for reliable and trustworthy approaches in checking the veracity of information in contemporary society.

One of the most important characteristics of new parliamentary data is its direct correspondence to the most recent events, including the ones with global impact on human health, social life and economics such as the current COVID-19 pandemic. By comparing the data synchronically and diachronically within a cross-lingual context, scientific and civil communities will be able to track pan-European discussion and can be quickly updated on any emerging topic.

Mission and Goal

The mission of the ParlaMint project is to turn existing contemporary multilingual and diverse cross-national parliamentary data into resources that are:



<https://www.clarin.eu/content/parlamint-towards-comparable-parliamentary-corpora>

SSHOC

EU Horizon 2020 Project



Project:



SSHOC

social sciences & humanities open cloud



Horizon 2020
European Union Funding
for Research & Innovation

Type of action & funding:
Research and Innovation action
(INFRAEOSC-04-2018)

Partners: 45

(20 beneficiaries + 25 LTPs)

SSH ESFRI Landmarks and Projects
& international SSH data infrastructures

Project budget:
€ 14,455,594.08

Duration: 40 months
(January 2019 – 30 April 2022)

Project website:
www.SSHopencloud.eu



Objectives:

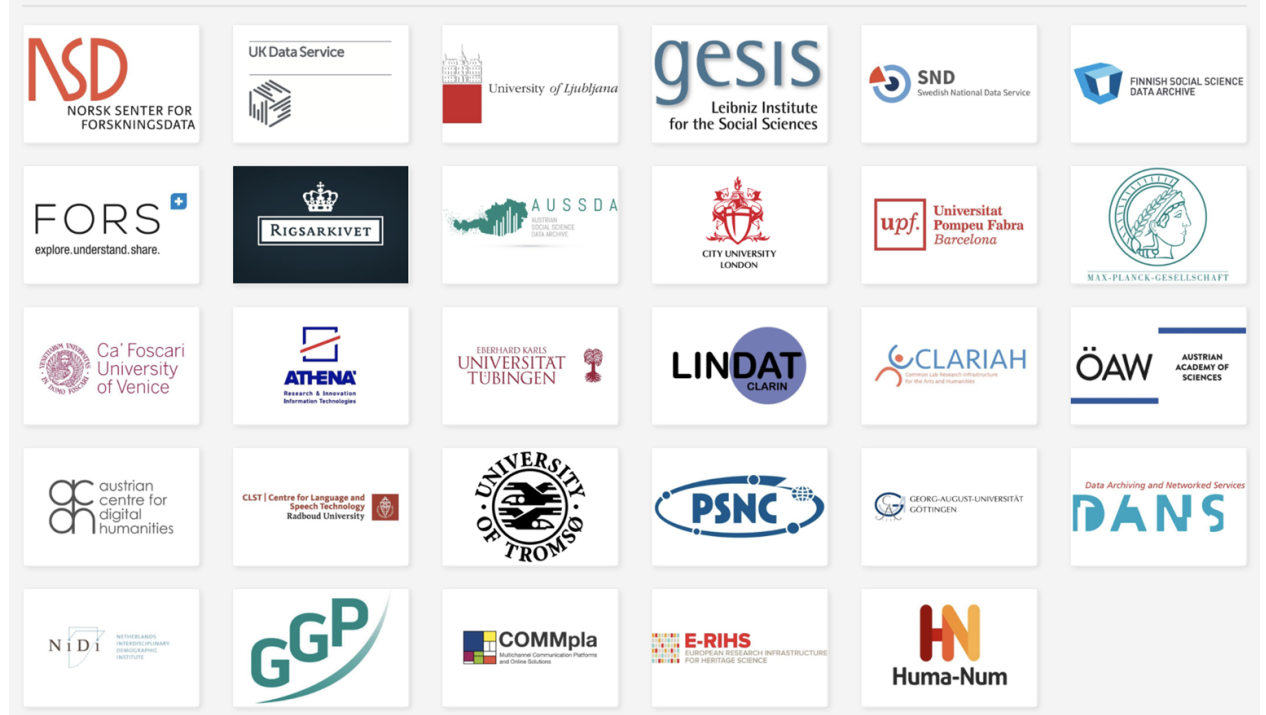
- creating the social sciences and humanities (**SSH**) part of European Open Science Cloud (**EOSC**)
- maximising **re-use** through **Open Science** and **FAIR** principles (standards, common catalogue, access control, semantic techniques, training)
- interconnecting existing and new infrastructures (clustered cloud infrastructure)
- establishing appropriate **governance model** for SSH-EOSC

Partners

[Home](#) / [Partners](#)



Collaborating Organisations



27 + 29 > 45!

Expected impact



The Social Sciences and Humanities are seamlessly integrated in the European Open Science Cloud



Availability of an EU-wide, easy-to-use SSH Open Marketplace, where tools and data are openly accessible



EU-wide availability of high quality "cloud ready" SSH tools and high quality SSH data



EU-wide availability of trusted and secure access mechanisms for SSH data, conforming to EU legal requirements

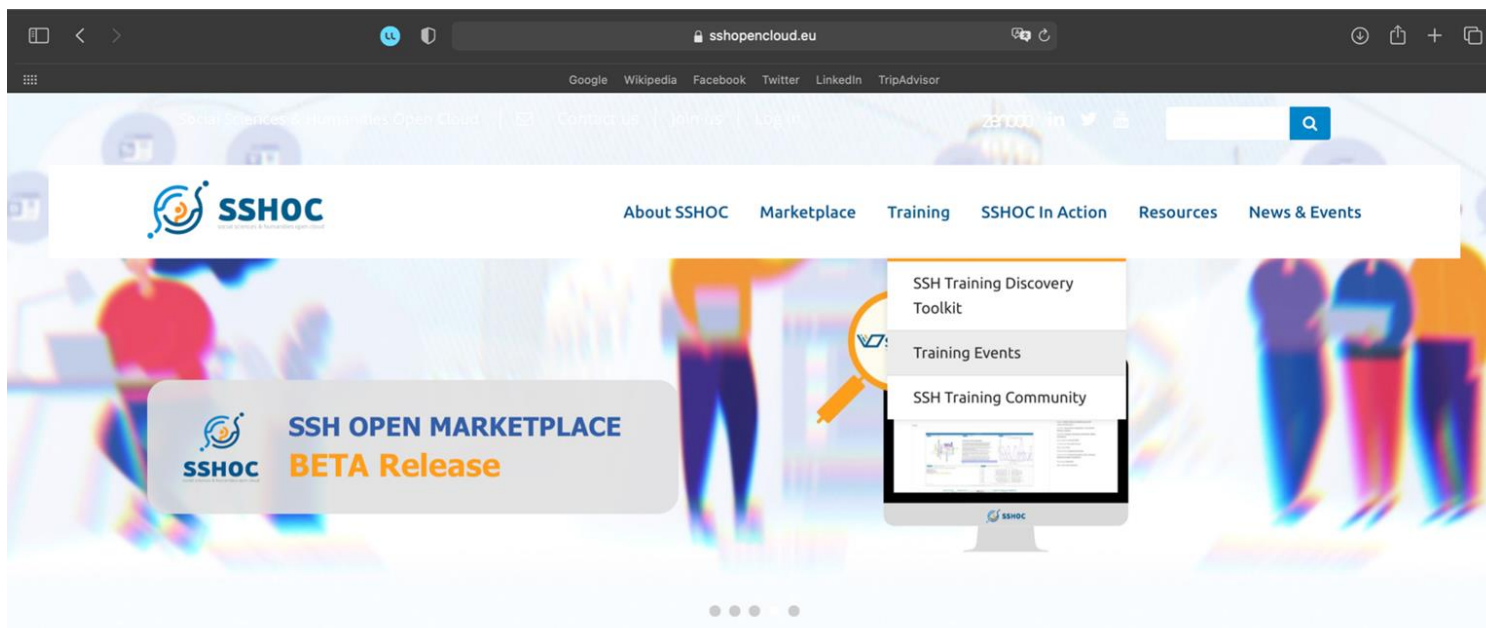


State of the art Research Infrastructure in several pilot domains advanced through dedicated SSH data pilots cluster projects



Maximising reuse through Open Science and FAIR principles (standards, common catalogue, access control, semantic techniques, training)

SSHOC website: past and future events (and relevant materials)



SSHOC Principal Goals

-  SSHOC will create the social sciences and humanities area of the **European Open Science Cloud (EOSC)** thereby facilitating access to flexible, scalable research data and related services streamlined to the precise needs of the SSH community.
-  SSHOC will leverage and interconnect existing and new infrastructures from the **SSH ERICs** to foster synergies across disciplines and expedite interdisciplinary research and collaboration.
-  To maximise the efficiency and effectiveness of data re-use **SSHOC** will apply Open Science practices and **FAIR** principles to data management.
-  **SSHOC** will set up an appropriate governance model for the social sciences and humanities area of the **EOSC**, taking into account the specificities of different sub-domains within

<https://sshopencloud.eu/training/training-events>

<https://sshopencloud.eu/events/sshoc-workshop-data-citation-practice>



SSHOC website

[About SSHOC](#)[Marketplace](#)[Training](#)[SSHOC In Action](#)[Resources](#)[News & Events](#)

Search

[Home](#) / [Search](#)[Content](#)[Users](#)

Enter your keywords

Search results

- [SSHOC Report on \(meta\) data interoperability problems: What research librarians need to know](#)
... 02 December 2019 Are you a **librarian** with responsibility for data management? If so, you may gain valuable ...
Anonymous (not verified) - 01/10/2020 - 12:40
- [SSHOC-DARIAH Train-the-Trainer Research Data Management Bootcamp](#)
... some other cases, we see the emergence of data steward, data **librarian**, Open Science officer roles (either full-or part-time or on ...
[i.mazourine](#) - 17/03/2021 - 09:33

News

[SSHOC Takes Over @CLARINERIC Twitter Account](#)

Between 10:00 and 13:00 on 21 June, SSHOC will take over the Twitter account...

[ESFRI Science Clusters Position Statement on Expectations and Long-Term Commitment in Open Science](#)

The Science Cluster (ENVRI-FAIR, EOSC-Life, ESCAPE, PANOSC and SSHOC) delivered...

[SSHOC Workshop Notes: Citizen Science & Cultural Heritage. Planning for Success](#)

If you ever played with the idea of crowdsourcing your project, but you took a...



SSH Training Discovery Toolkit

The screenshot shows a web browser window with the URL `training-toolkit.sshopencloud.eu/entities?search=librarians`. The page header includes the SSHOC logo and navigation links for About, Privacy policy, Legal notice, and Contact. Below the header, a search bar contains the text 'librarians' and a 'Search' button. The main content area displays 'Search entities' and 'Displaying 1 - 9 of 9'. A table lists search results with columns for Source of item, Title, and Description. On the right side, there are filter panels for Entity type, Collections, and Organisation.

The SSH Training Discovery Toolkit provides an inventory of training materials relevant for the Social Sciences and Humanities. Use the search bar to discover materials or browse through the collections. The filters will help you identify your area of interest.

librarians

Search entities

Displaying 1 - 9 of 9

Source of item	Title	Description
Research Data Mantra	DIY Research Data Management Training Kit for Librarians	Training kit for librarians who wish to gain confidence and understanding of research data management, based on open educational materials, covering five topics:
DARIAH-CAMPUS	How to make the most of your publications in the humanities?	The workshop jointly organized by FOSTER Plus and DARIAH-EU will focus on domain-specific practices of opening up scholarly communication in Arts and Humanities.
Humanities Commons - Open access, open source, open to all	Making our Information Ecosystem Explicit	Although conversations about information literacy have grown substantially since the ACRL Competency Standards (2000) and the Framework for Information Literacy for Higher Education (2016) were int

Entity type

- Item (6)
- Source (3)

Collections

- Training Discovery Toolkit (8)

Organisation

- Consejo Superior de

SSH Open Marketplace

The screenshot shows the SSH Open Marketplace website. At the top left is the logo 'SSH Open Marketplace' with a 'BETA' badge. The navigation menu includes 'Tools & Services', 'Training Materials', 'Publications', 'Datasets', 'Workflows', 'Browse', and 'About'. There are also links for 'Report an issue' and 'Sign in'. The main heading is 'Social Sciences & Humanities Open Marketplace'. Below this, there is a paragraph: 'Discover new and contextualised resources for your research in Social Sciences and Humanities: tools, services, training materials, workflows and datasets. [Read more...](#)'. A second paragraph states: 'The SSH Open Marketplace is under development and the current content is subject to change. Final release is planned for December 2021.' Below the text is a search bar with a dropdown menu set to 'All categories', a search input field, and a 'Search' button. The background features an illustration of people interacting with digital screens and data visualizations. The bottom section is divided into 'Browse' and 'Last added' columns. The 'Browse' section has two sub-sections: 'Browse by activity' and 'Browse by keyword'. 'Browse by activity' lists categories like 'Analyzing (576)', 'Visual Analysis (303)', 'Content Analysis (226)', etc. 'Browse by keyword' lists terms like '3D (64)', 'Heritage Science (31)', 'OCR-HTR (22)', etc. The 'Last added' section features two items: 'Gephi' and 'Parlamin 1.0', each with a brief description and a 'Read more' link.

Discovery portal

with 3 key concepts in focus:

- Contextualisation
- Curation
- Community

Beta version:

marketplace.sshopencloud.eu/

SSH Open Marketplace: Training Materials

[Home](#) / [Search](#)

Search results (140)

Refine your search

[Clear filters](#)

Sort by name

◀ Previous 1 of 7 Next ▶

CATEGORIES

- | | |
|--|------|
| <input type="checkbox"/> Tools & Services | 1606 |
| <input checked="" type="checkbox"/> Training Materials | 140 |
| <input type="checkbox"/> Publications | 2986 |
| <input type="checkbox"/> Datasets | 2 |
| <input type="checkbox"/> Workflows | 29 |

ACTIVITIES

- | | |
|---|---|
| <input type="checkbox"/> Analyzing | 1 |
| <input type="checkbox"/> Content Analysis | 1 |
| <input type="checkbox"/> Contextualizing | 1 |
| <input type="checkbox"/> Interpreting | 1 |
| <input type="checkbox"/> Network Analysis | 1 |



2.1 Error rates and ground truth - Text Digitisation



Keywords: OCR HTR

No description provided.

[Read more](#)



3DHOP - How To



Keywords: 3D

No description provided.

[Read more](#)

SSH Open Marketplace: Workflows

CATEGORIES

- Tools & Services 20
- Training Materials 2
- Publications 56
- Workflows 4

ACTIVITIES

- Encoding 3
- Mapping 1
- Preservation Metadata 1

KEYWORDS

- TEI 1
- recommended 1

SOURCES

- SSK 4



Create a dictionary in TEI

Activities: Encoding
Keywords: TEI, recommended

This scenario sets out the best practices for creating a born-digital dictionary, especially with the TEI (Text Encoding Initiative). However, building a standardized lexicographical dataset is not only a data format problem, it is also an intellectual and technical process...

[Read more](#)



Creation of a TEI-based corpus

This scenario explains the steps to take, in order to create a corpus based on the TEI tagset. As of today, the TEI guidelines have become a de facto standard for text annotation, providing solutions for a great variety of text and phrase structures, information on content...

[Read more](#)



Creating interoperable TEI text resources with the DTA 'Base Format' (DTABf)

Activities: Encoding, Preservation Metadata

Currently, initiatives for the digitization of textual resources and their provision to the interested community are manifold and various. Hence, scholars who want to base their research on digitized texts, especially if working with popular works, may find a considerable...

[Read more](#)



Project-centered EAD customization

Activities: Encoding, Mapping

TEI ODD can be used to document data models external to the TEI environment. Several projects working with archival standards (in particular EAD) use it as well. PARTHENOS created and maintain an instance of the EAD specification in ODD, that can be used to create project...

[Read more](#)



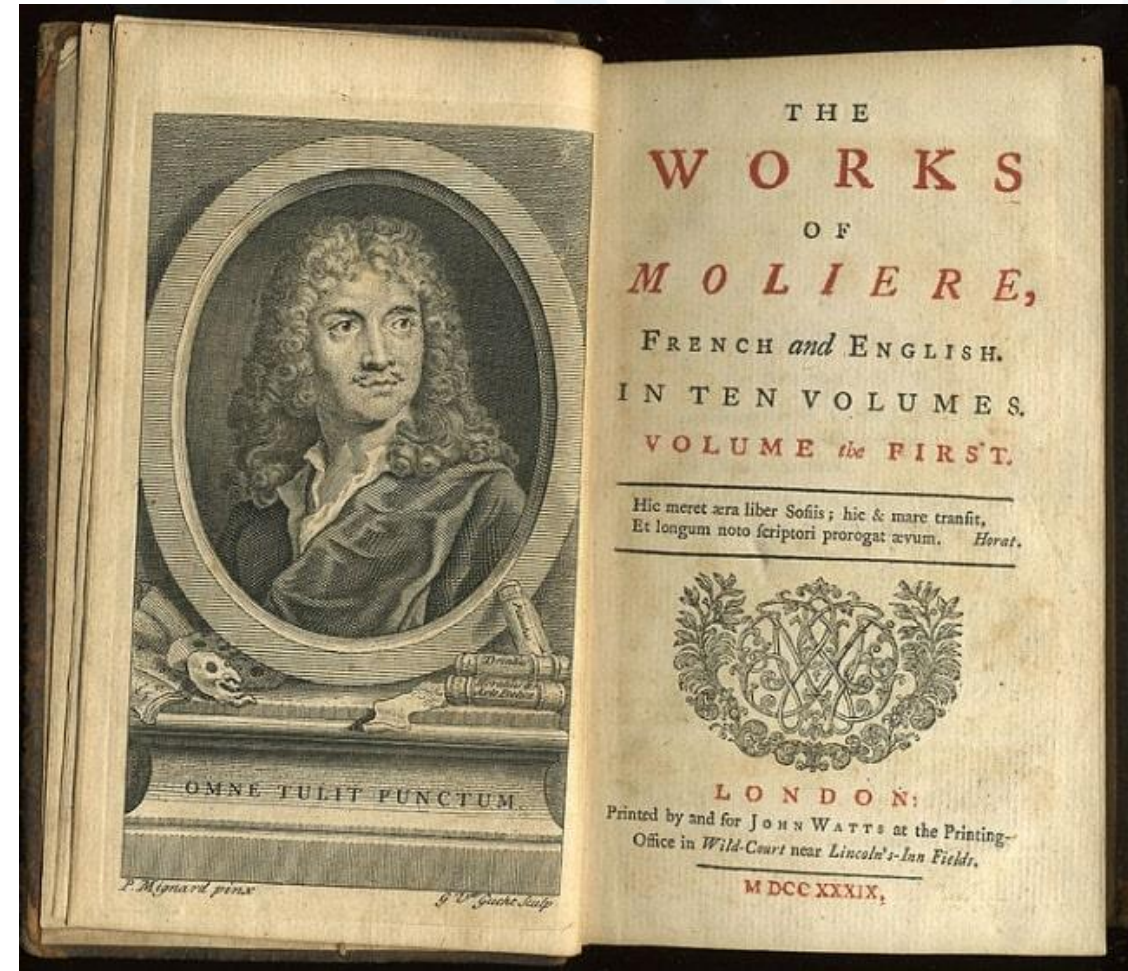
Scenario of use and motivation



Study of Theatrical Characters

Frontini, Francesca, Mohamed Amine Boukhaled, and Jean Gabriel Ganascia. 2018. '**Approaching French Theatrical Characters by Syntactical Analysis: A Study with Motifs and Correspondence Analysis**'. In Grammar of Genres and Styles. From Discrete to Non-Discrete Units, by Dominique Legallois, Thierry Charnois, and Meri Larjavaara, 320:118–39. Trends in Linguistics. Berlin/Boston: De Gruyter Mouton. <https://hal.archives-ouvertes.fr/hal-01832651>

Galleron, Ioana. 2016. '**Playing with French Drama: from Old Research Questions to New Research Tools**'. In Digital Humanities 2016: Conference Abstracts. Jagiellonian University & Pedagogical University, Kraków, pp. 522-523. <https://dh2016.adho.org/abstracts/289>



User: Researcher with Limited Knowledge of Digital Methods

Requiring access to:

- quality digital source data
 - richly encoded
 - metadata (author, year of publication, list of characters, ...)
 - annotation (lines of characters, dialogue, ...)
- easy to use tools to
 - explore
 - enrich
 - analyse

How can librarians help?

How can infrastructures help librarians?

Use the CLARIN Virtual Language observatory to find data

Virtual Language Observatory Search Contributors Help CLARIN

VLO / Faceted search / Search results

hamlet

Showing 1 to 10 of 401 results for hamlet Results per page: 10

Use the categories below to limit the search results to those matching the selected value(s).

- Language
- Collection
- Resource type
- Modality
- Format

Temporal Coverage

<< < 1 2 3 4 5 6 7 8 9 10 > >>

Hamlet
(Part of OTA Legacy Collection)
No description
English
Landing page for this record
The search results include 1 record with the same title.

Hamlet
(Part of TextGrid Repository)
No description
The search results include 2 records with the same title.

Access to Source Material



VLO / Faceted search / Search results / Record: Hamlet



Record 1 of 401



Hamlet



Record details

Links (5)

Availability

All metadata

Technical Details

Name	Type
 HDL 0121	landing page 
 dublin_core.xml	XML  
 metadata_local.xml	XML  
 header0121.xml	XML  
 hq11603-0121.txt	Plain Text  



Where does it come from?

Oxford Text Archive About OTA Electronic Enlightenment CLARIN

OTA Home / View Item Search

Hamlet

Please use the following text to cite this item or export to a predefined format: <http://hdl.handle.net/20.500.12024/1064>

Shakespeare, William, 1564-1616, *Hamlet*, Oxford Text Archive, <http://hdl.handle.net/20.500.12024/1064>.

Share: [f](#) [t](#) [g+](#)

Authors Shakespeare, William, 1564-1616

Date of publication 1609

Type Text

Language(s) English

OTA identifier ota:1064

Collection(s) Core Collection

[Show full item record](#)

This item is **Publicly Available** and licensed under:
Attribution-NonCommercial-ShareAlike 3.0 Unported (CC BY-NC-SA 3.0)

OXFORD TEXT ARCHIVE Bodleian Libraries UNIVERSITY OF OXFORD

Browse
> All of the Repository

My Account
Login

Statistics
Statistics **BETA**

General Information
Cite
Oxford University users
FAQ
About
Help Desk
Privacy policy



Process the text



Record 1 of 401


< previous next >



Hamlet

Record details Links (5) Availability All metadata Technical Details

Name	Type
 HDL 0121	landing page 
 dublin_core.xml	XML ... 
 metadata_local.xml	XML ... 
 header0121.xml	XML ... 
 hq11603-0121.txt	Plain Text ... 

 Process with Language Resource Switchboard

Annotate



↓ Process Input ↓

Output Text

Show Table

Show Trees

Save Output File

Id	Form	Lemma	UPosTag	XPosTag	Feats	Head	DepRel	Deps	Misc
# generator = UDPipe 2, https://lindat.mff.cuni.cz/services/udpipe									
# udpipe_model = english-ewt-ud-2.6-200830									
# udpipe_model_licence = CC BY-NC-SA									
# newdoc									
# newpar									
# sent_id = 1									
# text = The partners of my watch, bid them make haste.									
1	The	the	DET	DT	Definite=DeflPronType=Art	2	det	_	TokenRange=0:3
2	partners	partner	NOUN	NNS	Number=Plur	7	nsubj	_	TokenRange=4:12
3	of	of	ADP	IN	_	5	case	_	TokenRange=13:15
4	my	my	PRON	PRP\$	Number=SinglPerson=1lPoss=YeslPronType=Prs	5	nmod:poss	_	TokenRange=16:18
5	watch	watch	NOUN	NN	Number=Sing	2	nmod	_	SpaceAfter=No!TokenRange=19:24
6	,	,	PUNCT	,	_	7	punct	_	TokenRange=24:25
7	bid	bid	VERB	VB	Mood=ImplVerbForm=Fin	0	root	_	TokenRange=26:29
8	them	they	PRON	PRP	Case=AcclNumber=PlurlPerson=3lPronType=Prs	7	obj	_	TokenRange=30:34



Ham?



our coosin and dearest Sonne. <	ham	.> My lord, ti's not the
vs, go not to {Wittenberg.} <	ham	.> I shall in all my
Hamlet.} {Exeunt all but Hamlet.} <	ham	.> O that this too much
Hor.> Health to your Lordship. <	ham	.> I am very glad to
and your poore seruant euer. <	ham	.> O my good friend, I
Marcellus.} <Marc.> My good Lord. <	ham	.> I am very glad to
trowant disposition, my good Lord. <	ham	.> Nor shall you make mee
to see your fathers funerall. <	ham	.> O I pre thee do
Lord, it followed hard vpon. <	ham	.> Thrift, thrift, {Horatio}, the funerall
father. <Hor.> Where my Lord? <	ham	.> Why, im my mindes eye
he was a gallant King. <	ham	.> He was a man, take
Lord, the King your father. <	ham	.> Ha, ha, the King my
Gentlemen This wonder to you. <	ham	.> For Gods loue let me

Input format must match with tools to obtain good results

TEI in Details

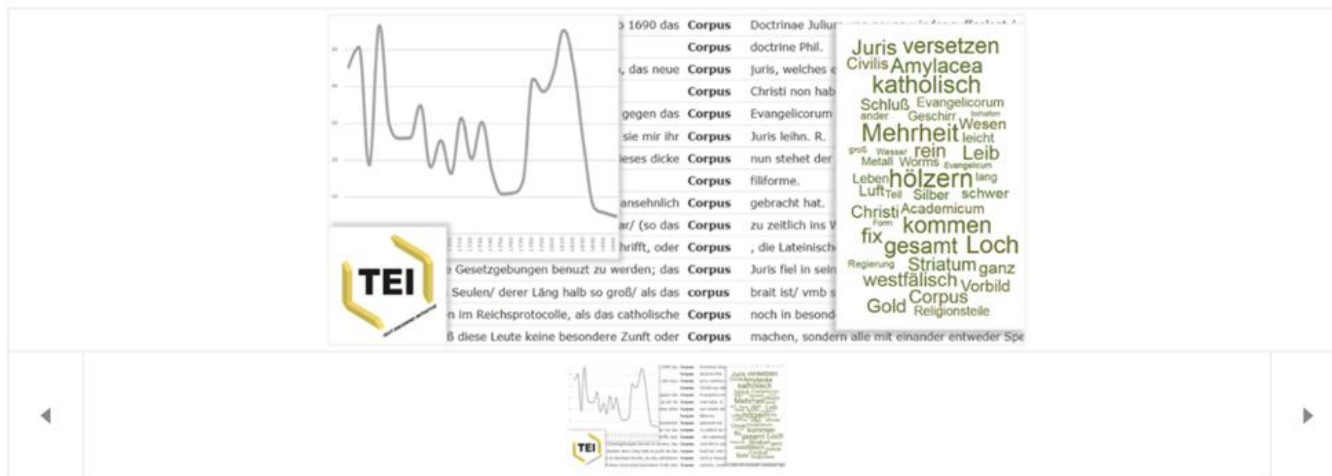


Existing SSHOC workflows



Creation of a TEI-based corpus

This scenario explains the steps to take, in order to create a corpus based on the TEI tagset. As of today, the TEI guidelines have become a de facto standard for text annotation, providing solutions for a great variety of text and phrase structures, information on content types, linguistic information on words or phrases, etc. In many digital text collections and digital edition projects annotation has been based on the TEI. Linguistic corpora based on TEI may thus be re-used in projects of other disciplines as well or may themselves benefit from the wide range of already existing resources.



Go to Workflow [↗](#)

Details

Contributors: Susanne Haaf, Klaus Illmayer, Piotr Bański

Language: eng

License: Creative Commons Attribution 4.0 International

Source: SSK

<https://marketplace.sshopencloud.eu/workflow/tEDt7j>

Corpora and collections in the VLO

english drama

Showing 1 to 10 of 496 results within selection for english drama x text/xml x

Results per page: 10

Use the categories below to limit the search results to those matching the selected value(s).

Language

Collection

Type to filter or search for more

Early English Books Online (Phase 1) (208)

OTA Core Collection (90)

OTA Legacy Collection (86)

Early English Books Online (Phase 2) (52)

Evans Early American Imprints (42)

ECCO - Eighteenth Century Collections Online (17)

CLARINO Bergen Centre (1)

Resource type

<< < 1 2 3 4 5 6 7 8 9 10 > >>

Records of early English drama : selections / compiled by Abigail Ann Young

(Part of OTA Core Collection)

No description

English English, Mid.. Latin Middle Engli..

Landing page for this record

10



Edward III (Drama)

(Part of OTA Legacy Collection)

No description

English

Landing page for this record

4





The search results include 1 record with the same title.




... and elsewhere

ABOUT ▾ CORPORA ▾ TOOLS ▾ HOW TO ▾ MERCH

Corpus	Number of plays	Number of characters	Text tokens	Stage Tokens	Last update	
Ger DraCor (German Drama Corpus)	537	12,920 (M: 9118, F: 2626)	9,478,256	375,316 (9,044,704)	177,863 (1,125,579)	11/05/2021, 11:09:12
Rus DraCor (Russian Drama Corpus)	211	3,647 (M: 2558, F: 861)	2,296,637	118,269 (2,173,058)	49,056 (212,960)	28/05/2021, 17:21:21
Ita DraCor (Italian Drama Corpus)	139	1,527 (M: 989, F: 413)	1,895,476	66,607 (1,763,669)	13,522 (62,311)	10/05/2021, 09:37:55
Swe DraCor (Swedish Drama Corpus)	68	769 (M: 382, F: 327)	737,001	35,420 (690,633)	17,209 (96,212)	14/12/2020, 18:29:38

If you want to cite DraCor, please use the following reference:

  Fischer, Frank, et al. (2019). Programmable Corpora: Introducing DraCor, an Infrastructure for the Research on European Drama. In *Proceedings of DH2019: "Complexities"*, Utrecht University, doi:10.5281/zenodo.4284002.

Drama Corpora Project 2020
Unless otherwise stated, all corpora and the web design are released under Creative Commons 0 1.0 
This site runs on  0.79.0 using  5.2.0

<https://dracor.org/>

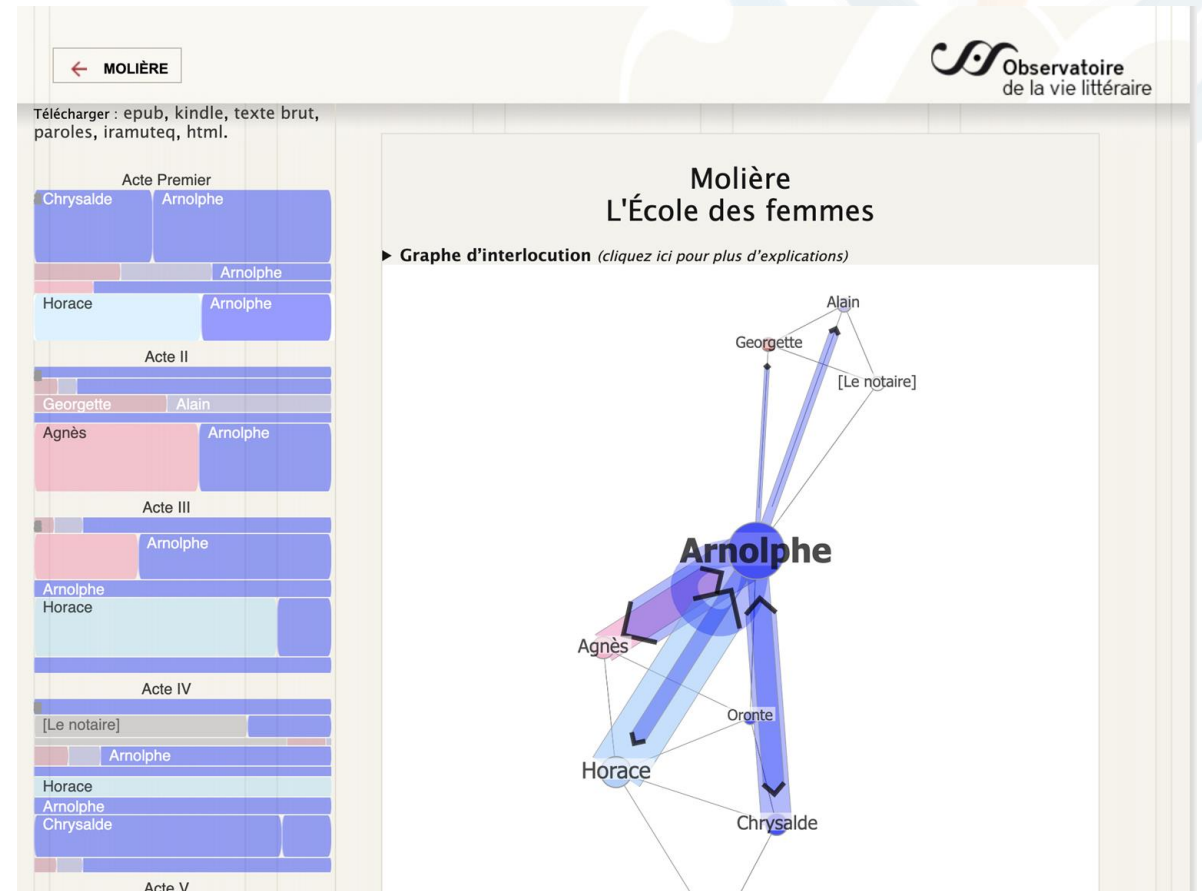
... and elsewhere

← MOLIÈRE, ACCUEIL

Observatoire de la vie littéraire

Théâtre de Molière

↓ Créé ↑	↓ Publié ↑	↓ Titre ↑
		La Jalousie du Barbouillé
		Le médecin volant
1670		Le Divertissement royal, mêlé de Comédie, de Musique, et d'Entrée de Ballet
1671		Livret de « Bourgeois gentilhomme », comédie-ballet
1671		Livret de « Psyché », tragicomédie-ballet
1682		Mélicerte
1655	1663	L'Étourdi ou Les Contre-temps
1656	1663	Le Dépit amoureux
1659	1660	Les Précieuses ridicules
1660	1660	Sganarelle ou Le Cocu imaginaire
1661	1661	L'École des maris
1661	1662	Les Fâcheux
1661	1682	Don Garcie de Navarre
1662	1663	L'École des femmes
1663	1663	La Critique de l'École des femmes
1663	1682	L'Impromptu de Versailles
1664	1664	La Princesse d'Élide
1664	1668	Le Mariage forcé
1664	1669	Le Tartuffe, ou L'Imposteur
1665	1666	L'Amour médecin
1665	1682	Don Juan, ou Le Festin de Pierre
1665	1683	Le Festin de Pierre
1666	1667	Le médecin malgré lui
1666	1667	Le Misanthrope
1667	1668	Le Sicilien
1668	1668	Amphitryon
1668	1669	L'Avare
1668	1669	George dandin
1669	1670	Monsieur De Pourceaugnac
1670	1671	Le Bourgeois gentilhomme
1670	1682	Les Amants magnifiques
1671	1671	Les Fourberies de Scapin
1671	1671	Psyché
1672	1672	Les femmes savantes



<https://obvil.sorbonne-universite.fr/corpus/molieres/molieres>

Relevant publications

[Home](#) / [Publications](#) / [To Catch a Protagonist - Quantitative Dominance Relations in German-Language Drama \(1730-1930\)](#)



To Catch a Protagonist - Quantitative Dominance Relations in German-Language Drama (1730-1930)

No description provided.

[Go to Publication](#) 

Details

Contributors: Frank Fischer 0005, Daniil Skorinkin, Peer Trilcke, Christopher Kittel, Carsten Milling

Pages: 193-200

Year: 2018

Source: DBLP





[\[English\]](#) [\[Deutsch\]](#) [\[Español\]](#) [\[Italiano\]](#) [\[Français\]](#) [\[日本語\]](#) [\[한국어\]](#) [\[中文\]](#)



Front Matter

- [Title](#)
- i. [Releases of the TEI Guidelines](#)
- ii. [Dedication](#)
- iii. [Preface and Acknowledgments](#)
- iv. [About These Guidelines](#)
- v. [A Gentle Introduction to XML](#)
- vi. [Languages and Character Sets](#)

Back Matter

- Appendix A [Model Classes](#)
- Appendix B [Attribute Classes](#)
- Appendix C [Elements](#)
- Appendix D [Attributes](#)
- Appendix E [Datatypes and Other Macros](#)
- Appendix F [Bibliography](#)
- Appendix G [Deprecations](#)
- Appendix H [Prefatory Notes](#)
- Appendix I [Colophon](#)

Text Body

- 1 [The TEI Infrastructure](#)
- 2 [The TEI Header](#)
- 3 [Elements Available in All TEI Documents](#)
- 4 [Default Text Structure](#)
- 5 [Characters, Glyphs, and Writing Modes](#)
- 6 [Verse](#)
- 7 [Performance Texts](#)
- 8 [Transcriptions of Speech](#)
- 9 [Dictionaries](#)
- 10 [Manuscript Description](#)
- 11 [Representation of Primary Sources](#)
- 12 [Critical Apparatus](#)
- 13 [Names, Dates, People, and Places](#)
- 14 [Tables, Formulae, Graphics and Notated Music](#)
- 15 [Language Corpora](#)
- 16 [Linking, Segmentation, and Alignment](#)
- 17 [Simple Analytic Mechanisms](#)
- 18 [Feature Structures](#)
- 19 [Graphs, Networks, and Trees](#)
- 20 [Non-hierarchical Structures](#)
- 21 [Certainty, Precision, and Responsibility](#)
- 22 [Documentation Elements](#)
- 23 [Using the TEI](#)

TEI sourcecode

- [Getting and Using the TEI Sources.](#)
- [TEI GitHub Repository](#)
- [Bug Reports, Feature Requests, etc.](#)

[\[English\]](#) [\[Deutsch\]](#) [\[Español\]](#) [\[Italiano\]](#) [\[Français\]](#) [\[日本語\]](#) [\[한국어\]](#) [\[中文\]](#)



TEI Default Text Structure

```
<TEI xmlns="http://www.tei-c.org/ns/1.0">
  <teiHeader>
    <!-- ... -->
  </teiHeader>
  <text>
    <front>
      <!-- front matter of copy text, if any, goes here -->
    </front>
    <body>
      <!-- body of copy text goes here -->
    </body>
    <back>
      <!-- back matter of copy text, if any, goes here -->
    </back>
  </text>
</TEI>
```



TEI Header. Minimal and recommended headers

```
<teiHeader>
  <fileDesc>
    <titleStmt>
      <title>Thomas Paine: Common sense, a
        machine-readable transcript</title>
      <respStmt>
        <resp>compiled by</resp>
        <name>Jon K Adams</name>
      </respStmt>
    </titleStmt>
    <publicationStmt>
      <istributor>Oxford Text Archive</istributor>
    </publicationStmt>
    <sourceDesc>
      <bibl>The complete writings of Thomas Paine, collected and edited
        by Phillip S. Foner (New York, Citadel Press, 1945)</bibl>
    </sourceDesc>
  </fileDesc>
</teiHeader>
```



TEI Text Body. Names, dates, people, and places

That silly man `<name role="politician" type="person">David Paul Brown</name>`
has suffered ...

I never fly from `<name key="LHR" type="place">Heathrow Airport</name>` to
`<name key="FR" type="place">France</name>`

`<date when="1807-06-09">June 9th</date>` The period is approaching which will terminate my present
copartnership. On the `<date when="1808-01-01">1st Jany.</date>` next, it expires by its own limitation.

`<placeName period="#christian">Stauropolis</placeName>`



To learn more about TEI ...

- TEI guidelines
 - <https://tei-c.org/release/doc/tei-p5-doc/en/html/index.html>
- TEI Tutorials (DARIAH TEACH)
 - Marjorie Burghart and Elena Pierazzo (2017). Digital Scholarly Editions: Manuscripts, Texts and TEI Encoding. DARIAH Teach. [Training module]. <https://teach.dariah.eu/course/view.php?id=32>
 - Laurent Romary (2020). The TEI Guidelines: Born to be Open. Edited by Maria Wiederänders. Austrian Centre for Digital Humanities and Cultural Heritage (ACDH-CH). [Video]. <https://youtu.be/hV-wtGlx8I8>
 - Toma Tasovac (2016). Digitising Dictionaries. DARIAH Teach. [Training module]. <https://teach.dariah.eu/course/view.php?id=20>



Q&A



Coffee-break

(5 minutes)



SSHOC/CLARIN use case

developed by

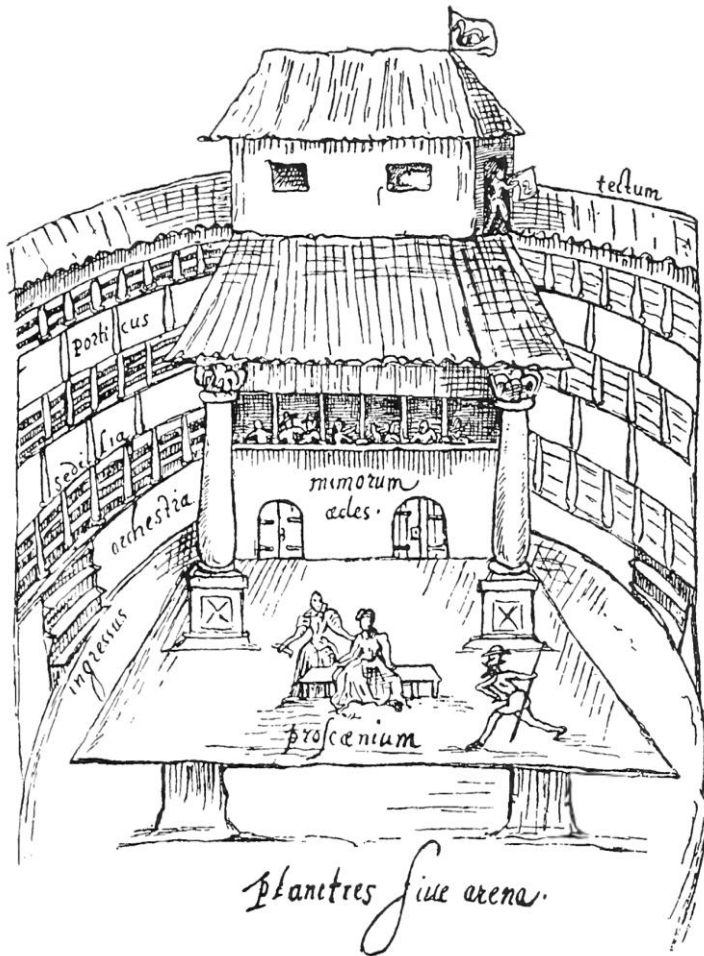
*DARIAH/UGOE (Georg-August-Universität Göttingen)
and*

CLARIN/LINDAT (Charles University of Prague)



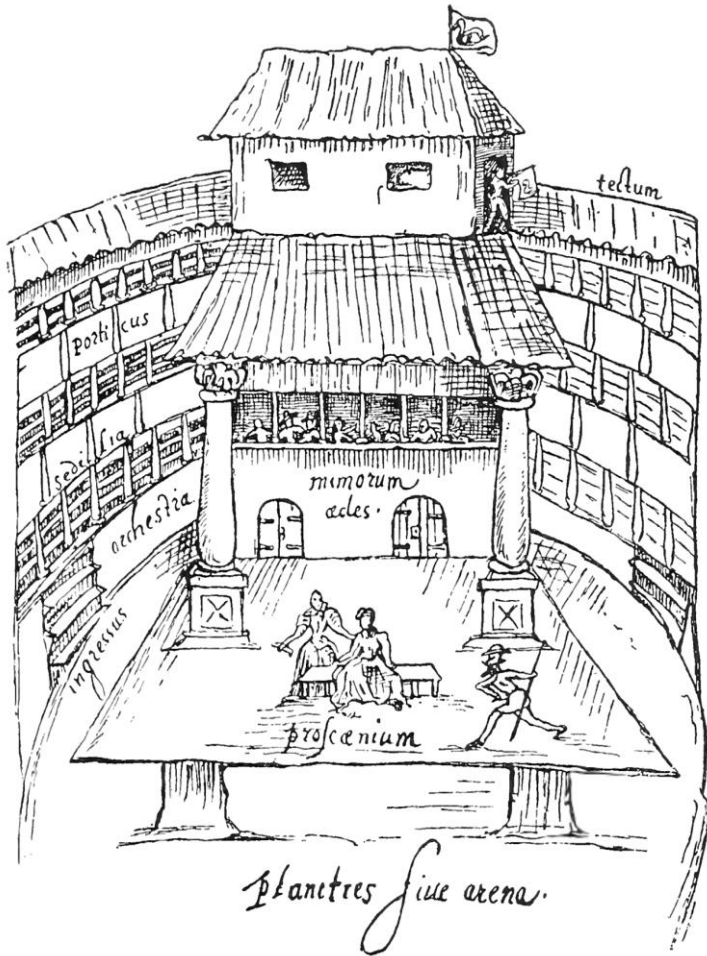
Intertextuality phenomena in European drama history.

Research context



- **When?**
 - 16th-17th century - the most productive time in the history of theatre -> **Large body of texts in English, French, Spanish.**
- **What?**
 - “Servant” character or other representatives of lower social classes have the function to **unveil the hidden order of discourse** by incorporating a comical wisdom which can be solely represented by minor characters.
- **Why is it interesting?**
 - Comparative literary analysis
 - Analysis of the literary language of individual dramas of the respective historical language level.

Intertextuality phenomena in European drama history. Challenges for a SSH scholar and RQ reformulation



- **Challenges for a SSH scholar**
 - Volume of material - not feasible to do it manually
 - Multilinguality
 - Absence of annotation of such characters
- **Research question as formulated in DH context:**
Can the relationship between characters be quantitatively recorded algorithmically, and is there a recurring pattern discernible?

Data

Language	Data	Encoding format	Link	Num files
English	Dramas from Shakespeare Library	TEI-XML-P5	dracor-org/shakedracor:Folger Shakespeare Library	37 plays
French	Théâtre Classique by Paul Fièvre	TEI P5	fredracor/tei	1452 files
Spanish	Dramas from Pedro Calderón de la Barca	partially in plain text that is to be converted to TEI	dracor-org/caldracor	54 dramas

Data analysis

- The main issue to address - **inconsistency of available formats**. Documents are sometimes:
 - available as TEI-XML, but not valid against any schema;
 - encoded with proprietary formats;
 - only available as plain text files.
- Approach
 - Different transformation XSLT and Python scripts for different parts of the corpus.
 - Itemization of parts of the corpus that can be taken into experiments.

Extraction of master's and servant's spoken text as single plain text files

William Shakespeare: »A Midsummer Night's Dream«ACT 1
Scene 1
Enter Theseus, Hippolyta, and Philostrate, with others.

THESEUS

Now, fair Hippolyta, our nuptial hour
Draws on apace. Four happy days bring in
Another moon. But, O, methinks how slow
This old moon wanes! She lingers my desires
Like to a stepdame or a dowager
Long withering out a young man's revenue.

HIPPOLYTA

Four days will quickly steep themselves in night;
Four nights will quickly dream away the time;
And then the moon, like to a silver bow
New-bent in heaven, shall behold the night
Of our solemnities.

THESEUS

Go, Philostrate,
Stir up the Athenian youth to merriments.
Awake the pert and nimble spirit of mirth.
Turn melancholy forth to funerals;
The pale companion is not for our pomp.
Philostrate exits.
Hippolyta, I wooed thee with my sword
And won thy love doing thee injuries,
But I will wed thee in another key,
With pomp, with triumph, and with reveling.

Enter Egeus and his daughter Hermia, and Lysander and Demetrius.

a-midsummer-night-s-dream.hippolyta_mnd.txt

Four days will quickly steep themselves in night
Four nights will quickly dream away the time
And then the moon like to a silver bow
New-bent in heaven shall behold the night
Of our solemnities
I was with Hercules and Cadmus once
When in a wood of Crete they bayed the bear
With hounds of Sparta Never did I hear
Such gallant chiding for besides the groves
The skies the fountains every region near
Seemed all one mutual cry I never heard

a-midsummer-night-s-dream.theseus_mnd.txt

Now fair Hippolyta our nuptial hour
Draws on apace Four happy days bring in
Another moon But O methinks how slow
This old moon wanes She lingers my desires
Like to a stepdame or a dowager
Long withering out a young man's revenue
Go Philostrate
Stir up the Athenian youth to merriments
Awake the pert and nimble spirit of mirth
Turn melancholy forth to funerals
The pale companion is not for our pomp
Hippolyta I wooed thee with my sword
And won thy love doing thee injuries
But I will wed thee in another key
With pomp with triumph and with reveling
Thanks good Egeus What's the news with thee
What say you Hermia Be advised fair maid
To you your father should be as a god
One that composed your beauties yea and one
To whom you are but as a form in wax
By him imprinted and within his power
To leave the figure or disfigure it

Extraction of master's and servant's spoken text as single plain text files

```
▼<?xml-model
  http://www.tei-c.org/release/xml/tei/custom/schema/relaxng/tei_all.rng application/xml http://relaxng.org/ns/structure/1.0
?>
▼<?xml-model
  http://www.tei-c.org/release/xml/tei/custom/schema/relaxng/tei_all.rng application/xml http://purl.oclc.org/dsdl/schematron
?>
▼<TEI xmlns="http://www.tei-c.org/ns/1.0">
  ▼<teiHeader>
    ▼<fileDesc>
      ▼<titleStmnt>
        <title>A Midsummer Night's Dream</title>
        <author key="wikidata:Q692">William Shakespeare</author>
        <editor xml:id="BAM">Barbara A. Mowat</editor>
        <editor xml:id="PW">Paul Werstine</editor>
      ▼<respStmnt>
        <resp>Encoded in TEI Simple by</resp>
        <name xml:id="MM">Martin Mueller</name>
        <name xml:id="MSP">Michael Poston</name>
      </respStmnt>
      ▼<respStmnt>
        <resp>Linguistically annotated with MorphAdorner by</resp>
        <name>Philip R. Burns</name>
        <name>Martin Mueller</name>
      </respStmnt>
    </titleStmnt>
  ▼<editionStmnt>
    ▼<edition n="0.5">
      An early release. Some encoding choices remain to be refined or extended.
    </edition>
  </editionStmnt>
  ▼<publicationStmnt>
    <publisher>Folger Digital Texts</publisher>
    <idno>MND</idno>
  ▼<address>
    <addrLine>http://www.folgerdigitaltexts.org</addrLine>
  </address>
  ▼<availability>
    ▼<licence target="http://creativecommons.org/licenses/by-nc/3.0/deed.en_US">
      Distributed under a Creative Commons Attribution-NonCommercial 3.0 Unported License
    </licence>
  </availability>
  <date>February, 2017</date>
  <idno type="wikidata" xml:base="https://www.wikidata.org/wiki/">Q104871</idno>
  <idno type="dracor" xml:base="https://dracor.org/id/">shake000008</idno>
</publicationStmnt>
  ▼<sourceDesc>
  ▼<biblFull>
    ▼<titleStmnt>
      <title>A Midsummer Night's Dream</title>
      <author>William Shakespeare</author>
      <editor>Barbara A. Mowat</editor>
      <editor>Paul Werstine</editor>
```

```
▼<body>
  ▼<div type="act" n="1">
    ▼<head>
      <w xml:id="fs-mnd-000010">ACT</w>
      <c></c>
      <w xml:id="fs-mnd-000030">1</w>
    </head>
  ▼<div type="scene" n="1">
    ▼<head>
      <w xml:id="fs-mnd-000040">Scene</w>
      <c></c>
      <w xml:id="fs-mnd-000060">1</w>
    </head>
  ▼<stage xml:id="stg-0000" n="SD 1.1.0" type="entrance" who="#Theseus_MND #Hippolyta_MND #Philostrate_MND #ATTENDANTS_MND">
    <w xml:id="fs-mnd-000070" n="SD 1.1.0">Enter</w>
    <c></c>
    <w xml:id="fs-mnd-000090" n="SD 1.1.0">Theseus</w>
    <pc xml:id="fs-mnd-000100" n="SD 1.1.0">,</pc>
    <c></c>
    <w xml:id="fs-mnd-000120" n="SD 1.1.0">Hippolyta</w>
    <pc xml:id="fs-mnd-000130" n="SD 1.1.0">,</pc>
    <c></c>
    <w xml:id="fs-mnd-000150" n="SD 1.1.0">and</w>
    <c></c>
    <w xml:id="fs-mnd-000170" n="SD 1.1.0">Philostrate</w>
    <pc xml:id="fs-mnd-000180" n="SD 1.1.0">,</pc>
    <c></c>
    <w xml:id="fs-mnd-000200" n="SD 1.1.0">with</w>
    <c></c>
    <w xml:id="fs-mnd-000220" n="SD 1.1.0">others</w>
    <pc xml:id="fs-mnd-000230" n="SD 1.1.0">.</pc>
  </stage>
  ▼<sp xml:id="sp-0001" who="#Theseus_MND">
  ▼<speaker xml:id="spk-0001">
    <w xml:id="fs-mnd-000240">THESEUS</w>
  </speaker>
  ▼<l xml:id="ftln-0001" n="1.1.1">
    <w xml:id="fs-mnd-000250" n="1.1.1" lemma="now" ana="#av">Now</w>
    <pc xml:id="fs-mnd-000260" n="1.1.1">,</pc>
    <c></c>
    <w xml:id="fs-mnd-000280" n="1.1.1" lemma="fair" ana="#j">fair</w>
    <c></c>
    <w xml:id="fs-mnd-000300" n="1.1.1" lemma="Hippolyta" ana="#n1-nn">Hippolyta</w>
    <pc xml:id="fs-mnd-000310" n="1.1.1">,</pc>
    <c></c>
    <w xml:id="fs-mnd-000330" n="1.1.1" lemma="our" ana="#po">our</w>
    <c></c>
    <w xml:id="fs-mnd-000350" n="1.1.1" lemma="nuptial" ana="#j">nuptial</w>
    <c></c>
    <w xml:id="fs-mnd-000370" n="1.1.1" lemma="hour" ana="#n1">hour</w>
  </l>
```

...



Extraction of master's and servant's spoken text as single plain text files

1. Count the number of <person> in <listPerson> for each drama and store result in a separate plain text file.

a. **With Terminal:** `cp -rfv xsl/CountingPersonInListPerson.xsl tei/.; cd tei`

b. **With Oxygen XML Editor**

i. Run CountingPersonInListPerson.xsl in ~/testenvironment/english/tei.

ii. Create toc.txt (“table of contents”) of directory

›~/corpora/testenvironment/english/tei‹.

2. Move both CountingPersonInListPerson.xsl and CountingPersonInListPerson.txt out of ~/testenvironment/english/tei.

a. **With Terminal:** `sudo rm -rvf CountingPersonInListPerson.xsl; mv -fv`

`CountingPersonInListPerson.txt /;`

Extraction of master's and servant's spoken text as single plain text files

3. Turn on virtual Python environment
 - a. via source `../../..../virtualenvs/test1/bin/activate` or where else activate is located;
 - b. for Windows OS do cf. [12. Virtual Environments and Packages – Python 3.9.5 documentation](#).
4. For extraction of master's and servant's spoken text as single plain text files itself run
 - a.

```
for f in *.xml; do mkdir ${f%%.xml}; name=`echo ${f%%.xml} | sed 's#^.*/#g#`;
../../..../software/scriptsForAll/pyScripts/extract_speech.py --input-file $f --
output-prefix ${f%%.xml}/${name}; done
```
 - b. It creates in `~/testenvironment/english/tei` for every drama an eponymous subdirectory storing all files for single characters's spoken text into it.

Extraction of master's and servant's spoken text as single plain text files

5. Count the number of single files in every directory and store it into plain text file `countingNumberOfCharactersInEveryFile.txt` outside directory run
 - a. `for f in *.xml; do echo $f; ls ${f%%.xml} | wc -l; done > ../countingNumberOfCharactersInEveryFile.txt`
 6. Go with Terminal/Shell upwards to directory above.
 7. Re-structure every entry in `countingNumberOfCharactersInEveryFile.txt`
 - a. from
name of drama linebreak
intent Number of files in directory
 - a. to
name of drama tabulator number of files in directory
- Sort them, alphabetically, afterwards, ie. in Terminal:
- ```
python3.7 sortingContentOfPlainTextFileAlphabetically.py
```

# *Extraction of master's and servant's spoken text as single plain text files*

8. Erase all occurrences of `file:/Users/*username*/Documents/corpora/testenvironment/english/tei` and sort entries alphabetically
  - a. in Terminal: `python3.7 CountingPersonInListPerson.py`
9. To create a special list as in example below ( that is required for the check up) run a script that will store this output into `counted.txt`.
  - =IDENTISCH(B2; E2)
  - =IDENTISCH(B3; E3)
  - =IDENTISCH(B4; E4)
  - ...
  - =IDENTISCH(B204; E204)

a. in Terminal: `python3.7 countToAnyGivenNumber.py`



# *Extraction of master's and servant's spoken text as single plain text files*

10. Copy both lists from ›countingNumberOfCharactersInEveryFile.txt‹ and ›CountingPersonInListPerson.txt‹ into ›checkIfPyScriptSelectsAllCharactersCorrectly.xls‹.

In this way, values are compared for every drama with ›=IDENTISCH(B2; E2)‹ etc. in column f.

11. Create empty plain text file `checkIfPyScriptSelectsAllCharactersCorrectly.txt`.

a. in Terminal: `touch checkIfPyScriptSelectsAllCharactersCorrectly.txt`

and copy content of column ›F‹ into plain text file

›checkIfPyScriptSelectsAllCharactersCorrectly.txt‹.

# *Extraction of master's and servant's spoken text as single plain text files*

12. Run the check up script:

a. `python3.7 checkIfPyScriptSelectsAllCharactersCorrectly.py`

- number of ERROR(s) is counted,
- their index number in ›checkIfPyScriptSelectsAllCharactersCorrectly.xls‹ is shown,
- both is stored into ›checkIfPyScriptSelectsAllCharactersCorrectly\_done.txt‹.

# Steps to be taken by a SSH researcher



- Annotated data
  - Find via an aggregator, e.g. VLO, SSH Open Marketplace
  - Download from the original source of the data collection
- Workflows with scripts, processing examples, compatible tools
  - Find via SSH Open Marketplace, CLARIN LRS
- Data processing
  - Offline: following the instructions provided in a workflow
  - Online: LRS



# Q&A



# Wrap-up and useful references





# Take home messages

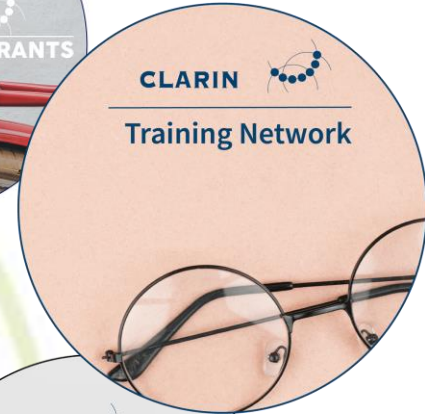
- **Collaboration with Research Infrastructures** is valuable for researchers and librarians who support researchers
- Tools, resources, services, and various teaching materials **can be found via aggregators**:
  - Broad SSH: SSH Open Marketplace
  - Language tools and resources for the SSH - Virtual Language Observatory
- The availability of standardised encoding for texts allows for **more advanced analysis**, and TEI in particular is a powerful annotation scheme for **information extraction**

# FURTHER SSHOC EVENTS AND WORKSHOPS

- **SSHOC at LIBER:**
  - **Wed, June 23, 11:00 - 12:30**, Session #3: Working with Software & Data Paper “Data citation for the Humanities and Social Sciences: a special case?” by Barbara McGillivray (University of Cambridge, UK), Nicolas Larrousse (TGIR Huma-Num, France), Daan Broeder (CLARIN ERIC)
  - Poster on the SSH Open Marketplace "Requesting Crowd Expertise: The SSH Open Marketplace and LIBER" will be presented by Stefan Buddenbohm at the LIBER poster session - **this Thursday 24 June from 12:30 - 13:00 CEST.**
- **Future SSHOC events - <https://www.sshopencloud.eu/sshoc-at-events>**
  - Workshop “SSHOC Vocabulary Initiative – What Users Want” at [ICTeSSH](#) on Monday, June 28, 2021, 9:00 AM – 10:30 AM
  - Workshop on Data protection and the GDPR - autumn 2021
  - Webinar on Data citation - autumn 2021

# CLARIN as ecosystem for knowledge exchange

- A network of **Knowledge Centres (K-centres)**
  - Help desk services
- **Sharing of expertise and best practices**
  - Annual Conference
  - Support for workshops and mobility
  - Ambassador network
- **Training** through life events, online courses and webinars
  - for developers
  - for end-users



# Getting involved in CLARIN. Channels

- Subscribe to CLARIN NewsFlash  
<https://www.clarin.eu/content/newsflash>
- Check out our past and futures events  
<https://www.clarin.eu/events>
- Open calls  
<https://www.clarin.eu/content/funding-opportunities>
- Follow us on Twitter @CLARINERIC

# Getting involved in CLARIN. Virtual events

- 11 CLARIN cafés - <https://www.clarin.eu/content/clarin-cafe>
- **First UPSKILLS Multiplier Event** “Every time I hire a linguist... Emergent tech profiles for linguists, translators and language experts” - 25 June, 10.00 - 16.00 CEST <https://www.clarin.eu/event/2021/first-upskills-multiplier-event>
- **ParlaMint Café** - 28th June 14:00-16:00 CEST  
<https://www.clarin.eu/event/2021/clarin-cafe-parlamint-unleashed>
- **CLARIN Annual Conference 2021** coming up - 27-29 September  
<https://www.clarin.eu/event/2021/clarin-annual-conference-2021-virtual-event>
- **Workshop “CLARIN and Libraries. Interoperability of Text Platforms for Digital Libraries”** - 10-12 and 14-16 CET, 15 October 2021

# Thank you for your attention!

*Questions?*

Please put them in the chat box.

Slides and a recording will be sent to all registered delegates.  
Post-event survey is shared with you via chat, and will be sent to you via email.  
<https://forms.gle/1WwxzkNxGL8qdjkL9>

Join our community



<https://www.sshopencloud.eu>



@SSHOpenCloud



[info@shopencloud.eu](mailto:info@shopencloud.eu)



[/in/shopencloud](https://www.linkedin.com/company/shopencloud)

