CLSINFRA

COMPUTATIONAL LITERARY STUDIES INFRASTRUCTURE

and Open Science

Maciej Eder (maciej.eder@ijp.pan.pl)

Bartłomiej Kunda (bartlomiej.kunda@ijp.pan.pl)

11.04.2024
Polish Open Science conference, Kraków, 10–12.04.2024

introduction

First, what CLS is about

- Computational Literary Studies
- Aimed at analyzing (large amounts of) textual data…
- … by computational techniques

Foundations of CLS

- Computation into criticism
- Distant reading
- Stylometry
- Authorship attribution
- Digital humanities
- Language resources
- Digital libraries
- Natural language processing
- Machine learning
- Big data
- ...

CLSINFRA

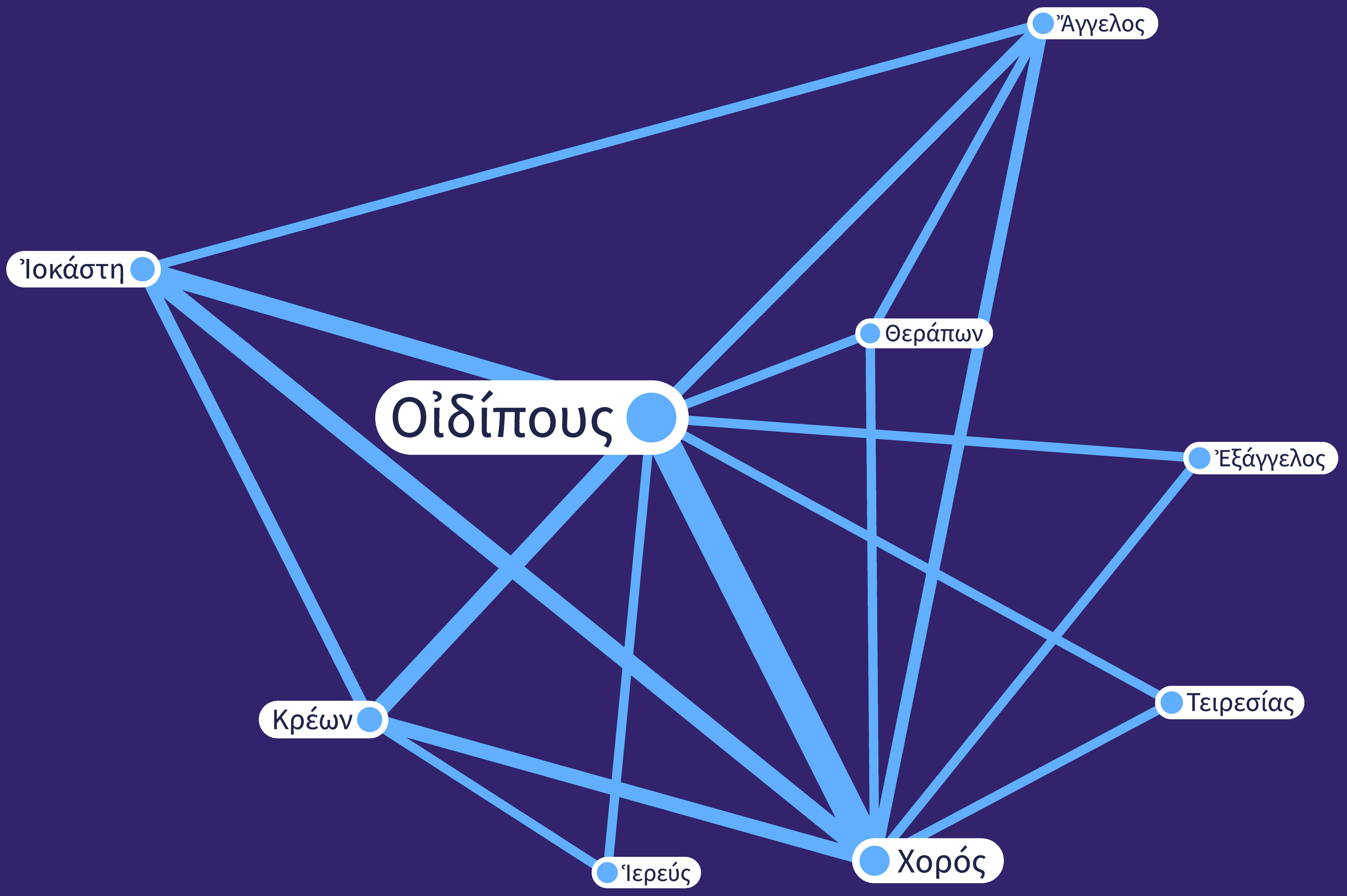
GreekDraCor

Oedipus Tyrannus

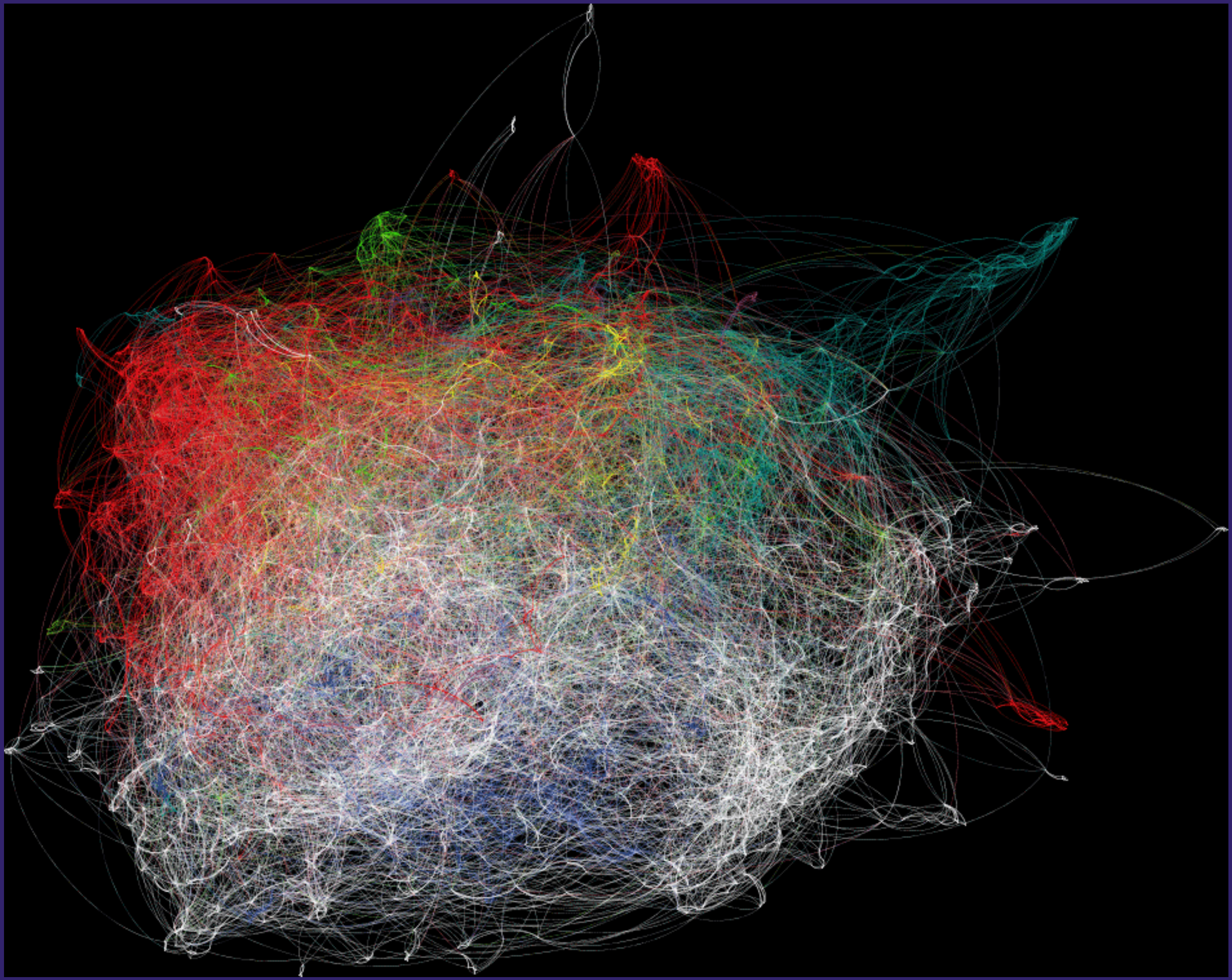
Ancient Greek tragedy

429 BCE

Sophocles



1,000 Polish novels



Combination of factors needed

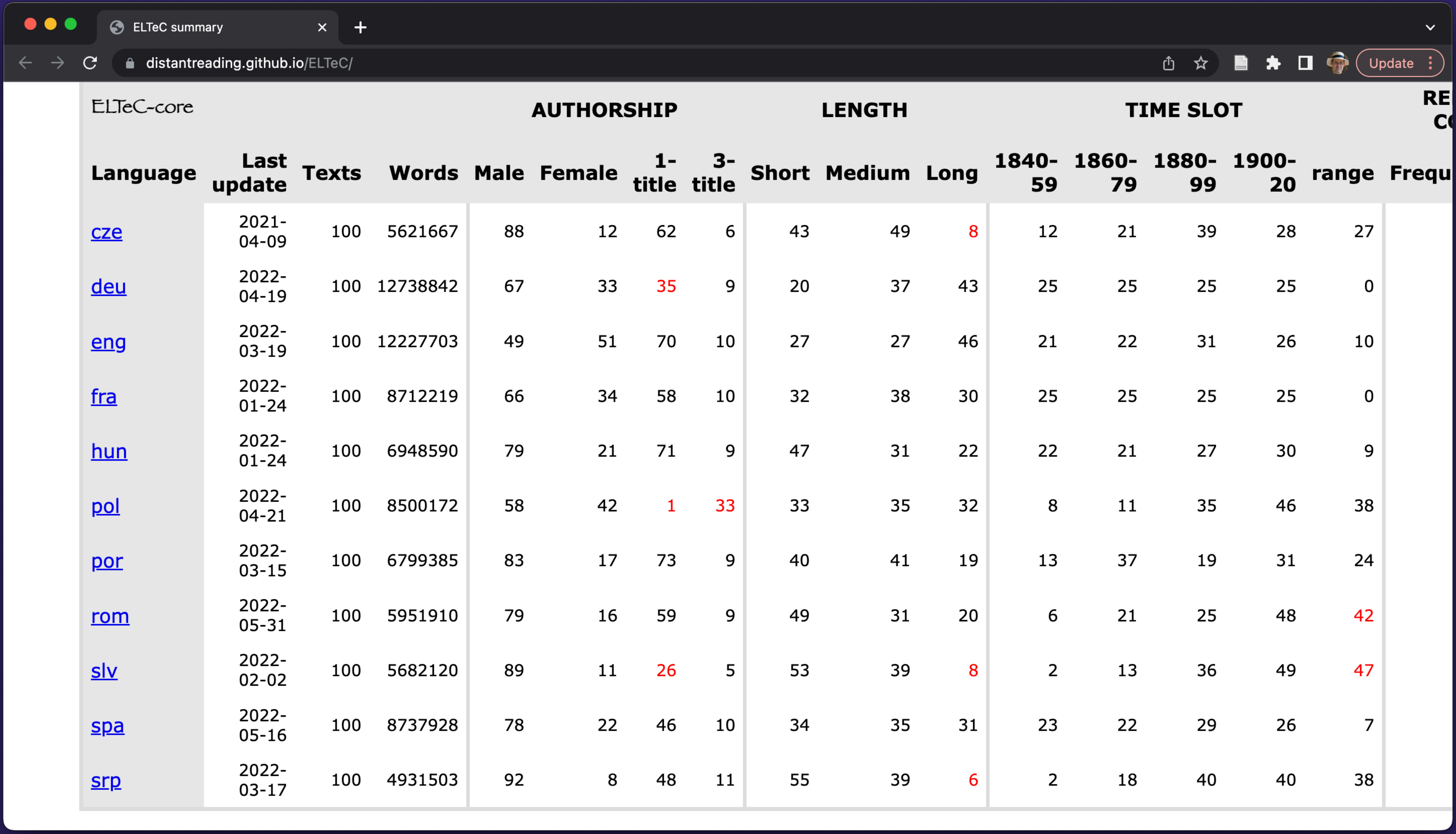
- Datasets (language resources)
- Tools (computer programs)
- Suitable methodology
- Computer power (i.e. scientific instruments)

Research infrastructures

infrastructures in DH

- in hard sciences, infrastructures are tangible
  - servers, telescopes, accelerators, …
- in the humanities, institutions are essential
  - libraries, publishing houses, journals, …
- in Digital Humanities, multifaceted needs
  - the notion of infrastructure needs reconsideration
  - corpora (FAIR!) but not only

ELTeC corpus



ELTeC-core				AUTHORSHIP				LENGTH			TIME SLOT					RE CO
Language	Last update	Texts	Words	Male	Female	1-title	3-title	Short	Medium	Long	1840-59	1860-79	1880-99	1900-20	range	Frequ
cze	2021-04-09	100	5621667	88	12	62	6	43	49	8	12	21	39	28	27	
deu	2022-04-19	100	12738842	67	33	35	9	20	37	43	25	25	25	25	0	
eng	2022-03-19	100	12227703	49	51	70	10	27	27	46	21	22	31	26	10	
fra	2022-01-24	100	8712219	66	34	58	10	32	38	30	25	25	25	25	0	
hun	2022-01-24	100	6948590	79	21	71	9	47	31	22	22	21	27	30	9	
pol	2022-04-21	100	8500172	58	42	1	33	33	35	32	8	11	35	46	38	
por	2022-03-15	100	6799385	83	17	73	9	40	41	19	13	37	19	31	24	
rom	2022-05-31	100	5951910	79	16	59	9	49	31	20	6	21	25	48	42	
slv	2022-02-02	100	5682120	89	11	26	5	53	39	8	2	13	36	49	47	
spa	2022-05-16	100	8737928	78	22	46	10	34	35	31	23	22	29	26	7	
srp	2022-03-17	100	4931503	92	8	48	11	55	39	6	2	18	40	40	38	

CLS INFRA

An infrastructural project for computational literary studies, founded by Horizon 2020 scheme

CLS INFRA project

- text collections (corpora)
  - quality
  - metadata
  - conversion
- methodology
  - tools (NLP, datavis, …)
  - tool chains
  - methodological considerations
  - bibliographic survey
- network of scholars
  - training schools
  - short-term research stays
  - collaboration with COST Action

Overarching idea is to connect…

- People
  - To establish a network of CLS researchers
- Data
  - To consolidate existing high-quality corpora…
  - …covering prose, drama and poetry
- Tools
  - To build a chain of NLP tools to analyze texts
- Methods
  - To provide a survey of state-of-the-art methods

activities

training schools

- Prague 2022
  - NLP tools
  - 25 participants on site
  - many more remotely
- Madrid 2023
  - text analysis
  - 10-11 May 2023
- Vienna 2024
  - corpus queries
  - 10-12 June 2024

TNA

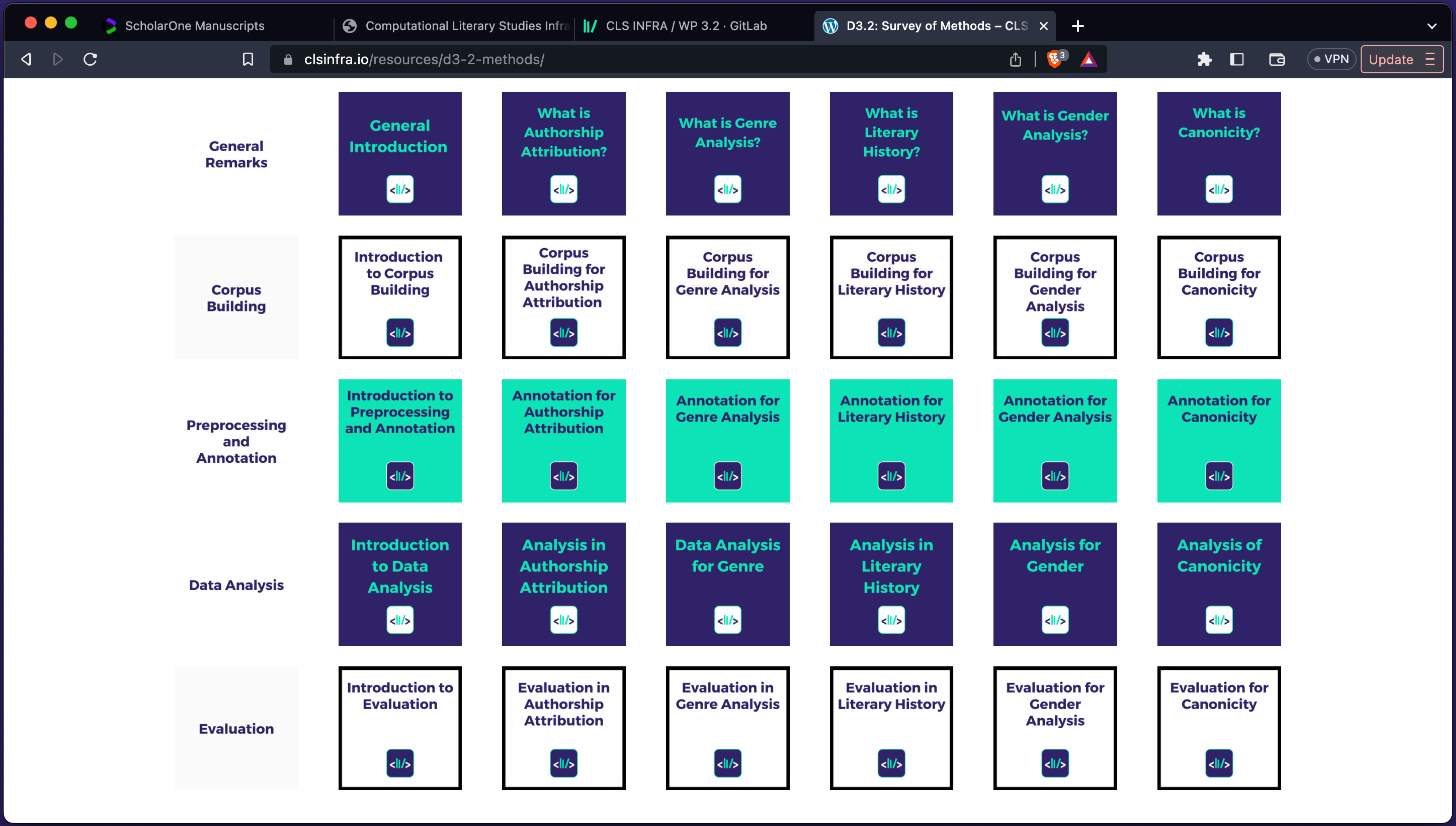
- transnational access
- short-term research stays…
- in one of 6 institutions:
  - NUI Galway
  - Uni Potsdam
  - Uni Trier
  - UNED Madrid
  - OEAW Vienna
  - Charles Uni, Prague
- everyone eligible
- two calls every year

CLS INFRA and Open Science

deliverables published

- 3.1 Report on the methodological baseline for (computational) literary studies
- 4.1 Report on the skills matrix for computational literary studies
- 5.1 Review of the data landscape
- 6.1 Assembly of existing data

survey of methods


ScholarOne Manuscripts
Computational Literary Studies Infra
CLS INFRA / WP 3.2 · GitLab
D3.2: Survey of Methods – CLS
clsinfra.io/resources/d3-2-methods/
VPN
Update

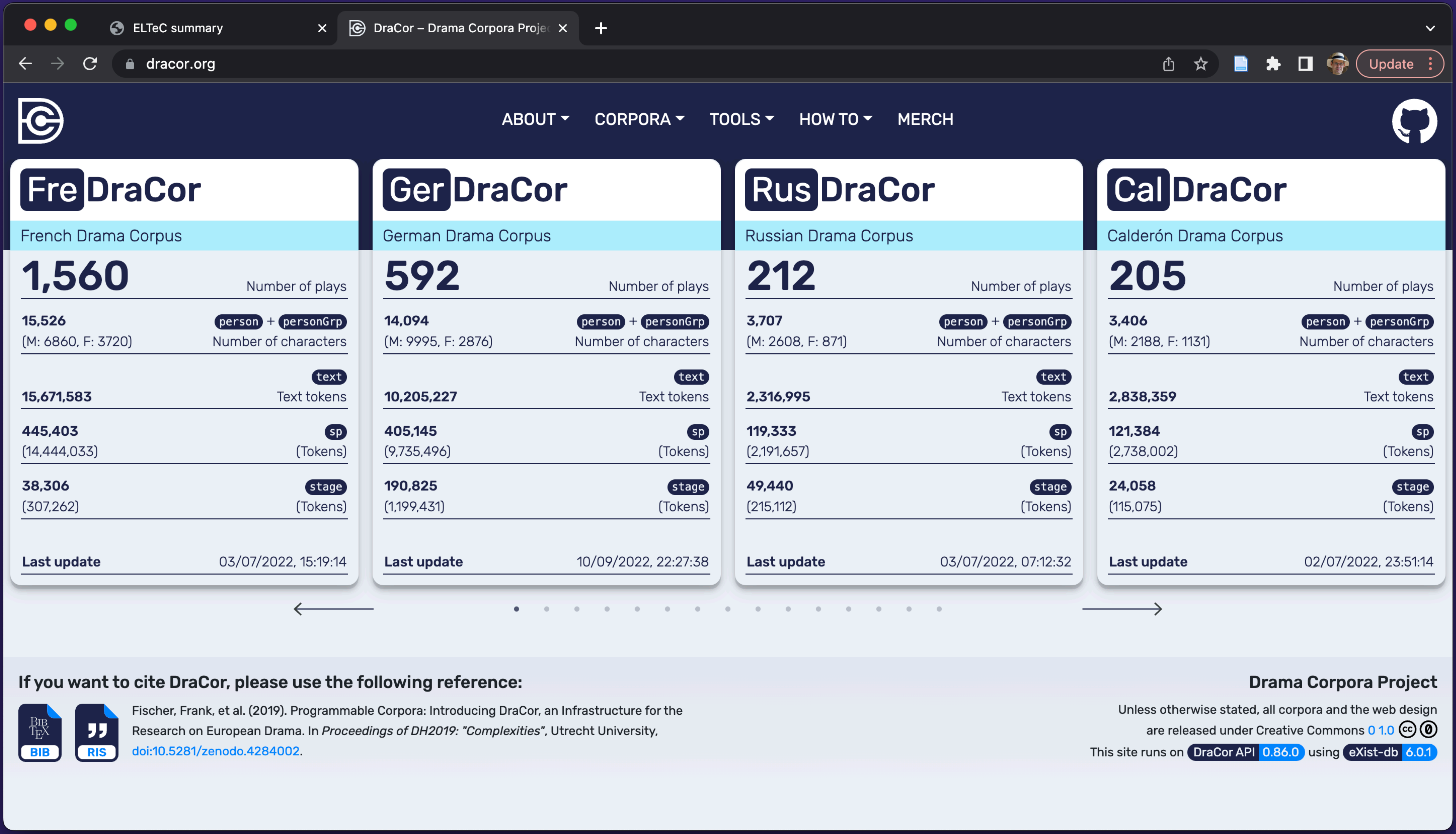
General Remarks	General Introduction	What is Authorship Attribution?	What is Genre Analysis?	What is Literary History?	What is Gender Analysis?	What is Canonicity?
Corpus Building	Introduction to Corpus Building	Corpus Building for Authorship Attribution	Corpus Building for Genre Analysis	Corpus Building for Literary History	Corpus Building for Gender Analysis	Corpus Building for Canonicity
Preprocessing and Annotation	Introduction to Preprocessing and Annotation	Annotation for Authorship Attribution	Annotation for Genre Analysis	Annotation for Literary History	Annotation for Gender Analysis	Annotation for Canonicity
Data Analysis	Introduction to Data Analysis	Analysis in Authorship Attribution	Data Analysis for Genre	Analysis in Literary History	Analysis for Gender	Analysis of Canonicity
Evaluation	Introduction to Evaluation	Evaluation in Authorship Attribution	Evaluation in Genre Analysis	Evaluation in Literary History	Evaluation for Gender Analysis	Evaluation for Canonicity

survey of methods

- Representation of the current CLS research based on a large collection of publications
- Introduction to CLS research areas and prominent issues
- A guide for further reading

https://methods.clsinfra.io/

programmable corpora: DraCor



FreDraCor

French Drama Corpus

1,560	Number of plays
15,526 (M: 6860, F: 3720)	person + personGrp Number of characters
15,671,583	text Text tokens
445,403 (14,444,033)	sp (Tokens)
38,306 (307,262)	stage (Tokens)
Last update	03/07/2022, 15:19:14

GerDraCor

German Drama Corpus

592	Number of plays
14,094 (M: 9995, F: 2876)	person + personGrp Number of characters
10,205,227	text Text tokens
405,145 (9,735,496)	sp (Tokens)
190,825 (1,199,431)	stage (Tokens)
Last update	10/09/2022, 22:27:38

RusDraCor

Russian Drama Corpus

212	Number of plays
3,707 (M: 2608, F: 871)	person + personGrp Number of characters
2,316,995	text Text tokens
119,333 (2,191,657)	sp (Tokens)
49,440 (215,112)	stage (Tokens)
Last update	03/07/2022, 07:12:32

CalDraCor

Calderón Drama Corpus

205	Number of plays
3,406 (M: 2188, F: 1131)	person + personGrp Number of characters
2,838,359	text Text tokens
121,384 (2,738,002)	sp (Tokens)
24,058 (115,075)	stage (Tokens)
Last update	02/07/2022, 23:51:14

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

BIB RIS

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

Unless otherwise stated, all corpora and the web design are released under Creative Commons 0 1.0

This site runs on DraCor API 0.86.0 using eXist-db 6.0.1

programmable corpora: DraCor

- DraCor: Highly functional prototype for an infrastructural ecosystem
- Programmable Corpora: research-driven API making text machine-actionable
- Open ecosystem: allowing for experimentation and discussion about architectural styles of research environments

https://dracor.org/

programmable corpora

- API Libraries developed in R and Python (published on the platforms PyPi an CRAN)
- Versioning:
  - Git commits for versioning and retrieving additional metadata
  - Docker containers of the entire research infrastructure - for more complex programmable corpora

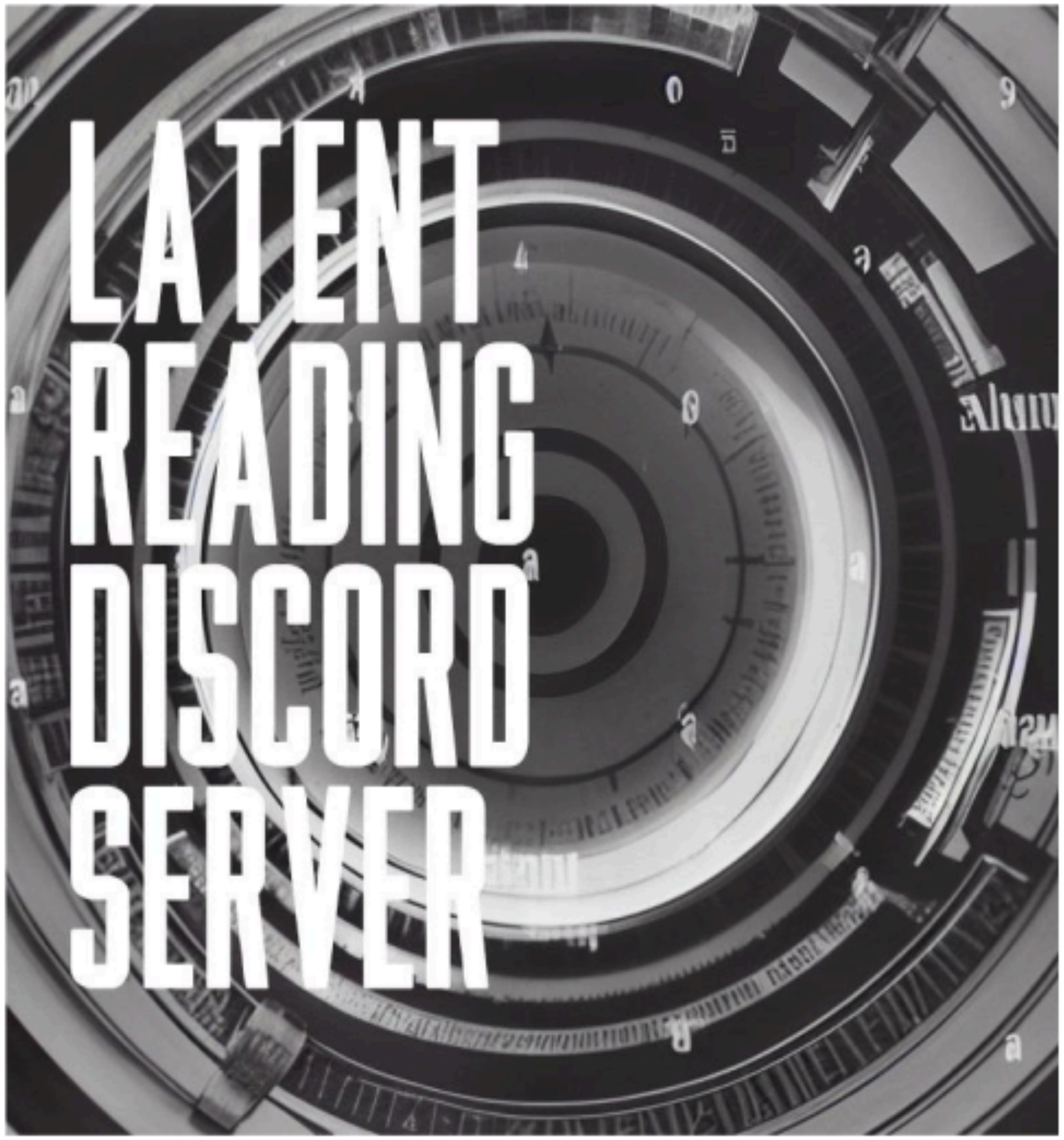
tools and access for CLS

- list and description of Natural Language Processing (NLP) tools (Corpus-Enrichment and NLP toolchain for common CLS research tasks)
- increasing the ease of access and application to NLP tools, as well as their standardization

CLS INFRA legacy

- all outputs available online (freely!)
- plans to develop the infrastructure
- a network of scholars is growing

CLS-centric Discord server



CLSINFRA COMPUTATIONAL LITERARY STUDIES INFRASTRUCTURE