

CLARIN - An Open Language Technology Research Infrastructure for SS&H Cluster

Maciej Piasecki

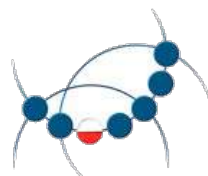
CLARIN-PL

Wrocław University of Science and Technology

CLARIN ERIC



CLARIN-PL
Common Language Resources and Technology Infrastructure



Wrocław University
of Science and Technology

Key points

- CLARIN ERIC – European Language Technology Research Infrastructure
- Involved in the EOSC SSH Open Cluster as one of the leading forces
- Powering EOSC SSH Open Cluster with CLARIN LT
- CLARIN distributed infrastructure of LT centres
- CLARIN-PL – a rich source of applications, tools and resources for SS&H
- Bottom-up collaboration in bringing large scale, distributed research environment to SS&H research community

CLARIN in countries and centres

Research Infrastructure CLARIN (= Common Language Resources and Technology Infrastructure)

A consortium of type ERIC (since 2012) - **Social Sciences and Humanities Cluster**

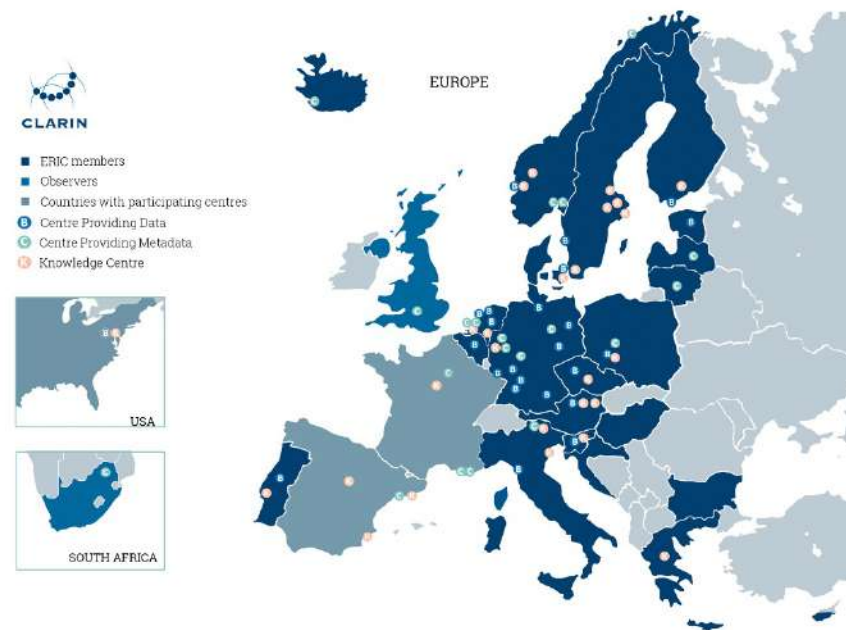
- 22 member countries
- 2 observers
- 1 linked party

A distributed network of >60 centres

25 CTS certified data centres,
strong focus on FAIRness & interoperability

- federated login: 
- central metadata harvesting for easy discovery: 
- chained services: 
- language data - in written, spoken, video or multimodal form
- advanced tools - to discover, explore, exploit, annotate, analyse or combine data sets, *wherever they are located*

Member of EOSC-Association  EOSC (since 2020)



Cluster collaboration of European RIs in SSH 1/2

Collaboration of all pan-European research infrastructures in SSH and their nodes

Support for research of cultural data, language data, survey data, and other relevant digital objects.


Objectives

- Federation of distributed SSH resources
- Enhanced methodological frameworks and workflows for the analysis of a.o.
 - multilingual data
 - multimedia data
 - heterogeneous data
 - mixed methods
- Increased potential for synergy and societal impact, within SSH and across clusters.




Cluster collaboration of European RIs in SSH 2/2

Common portal

- SSH Open Marketplace ([link](#))
 - faceted discovery platform
 - data, tools, training materials, publications, workflows
 - included in  eosoc catalogue



Societal impact agendas taking shape in the context of dedicated initiatives, such as

- Multidisciplinary collaboration in EOSC Future (H2020)  e.g. between
 - SSHOC and EOSC-Life (harmonised vocabularies for metadata)
 - SSHOC and ENVRI-Fair (climate neutral and smart cities)
- Challenge-driven funding schemes of Horizon Europe

Interconnecting existing and new infrastructures



CLARIN



European *Values* Study



Five European RI clusters

Aligned agendas and activities

- common web presence, white papers, etc.
- alignment of services to enable multi-disciplinary work
- projects aimed at common value proposition for ERA and EOSC

Example of cross-cluster project

- META-COVID
- Focus on vocabulary harmonisation across disciplinary domains of metadata for heterogeneous data resources related to COVID-19
- Applied to CLARIN's ParlaMint data, social media data, survey data.



CLARIN-PL w SSHOC (SSH Open Cloud) - EOSC (European Open Science Cloud)

The screenshot shows a web browser window displaying the B2DROP interface. The address bar shows the URL: <https://b2drop.eudat.eu/apps/files/?dir=/events/20181123-eosclaunch-vienna&fileid>. The page features a navigation bar with links for "WHAT IS B2DROP", "USER GUIDE", "FAQs", and "CONTACT". The main content area displays a file list for the directory "events > 20181123-eosclaunch-vienna".

Name	Size	Modified
male_female_speeches.zip	5.2 MB	3 days ago

The file details panel on the right shows the file name "male_female_speeches.zip" with a size of 5.2 MB and a modification time of 3 days ago. It includes options for "Activities", "B2SHARE", "Comments", "Sharing", and "Versions". The "Sharing" section is active, showing a checked "Share link" option and a generated link: <https://b2drop.eudat.eu/s/dFexm9tS7TJdoZE>. Other options include "Allow editing", "Password protect", and "Set expiration date".

Więcej: <https://www.clarin.eu/showcase/eosc-portal-demonstration>

CLARIN-PL in SSHOC (SSH Open Cloud) - EOSC (European Open Science Cloud)

The screenshot shows the B2DROP web interface. At the top, there is a navigation bar with the B2DROP and EUDAT logos, and links for 'WHAT IS B2DROP', 'USER GUIDE', 'FAQs', and 'CONTACT'. Below this is a file browser interface. On the left, there is a sidebar with navigation options: 'All files', 'Recent', 'Favorites', 'Shared with you', 'Shared with others', 'Shared by link', and 'Tags'. The main area shows a file named 'male_female_speeches.zip' with a size of 5.2 MB and a modification date of '3 days ago'. A context menu is open over the file, listing actions: 'Add to favorites', 'Details', 'Rename', 'Move or copy', 'Download', 'B2SHARE', 'Switchboard', and 'Delete'. On the right side, there is a detailed view of the file, including a share link: 'https://b2drop.eudat.eu/s/dFexm9tS7TjdoZE'. Below the link, there are checkboxes for 'Share link', 'Allow editing', 'Password protect', and 'Set expiration date'.

GO TO EUDAT WEBSITE

B2DROP **EUDAT**

WHAT IS B2DROP USER GUIDE FAQs CONTACT

All files Recent Favorites Shared with you Shared with others Shared by link Tags

events > 20181123-eosclaunch-vienna

Name	Size	Modified
male_female_speeches.zip	5.2 MB	3 days ago

- Add to favorites
- Details
- Rename
- Move or copy
- Download
- B2SHARE
- Switchboard
- Delete

male_female_speeches.zip

★ 5.2 MB, 3 days ago Tags

Activities B2SHARE Comments **Sharing** Versions

Name, federated cloud ID or email address...

Share link

https://b2drop.eudat.eu/s/dFexm9tS7TjdoZE

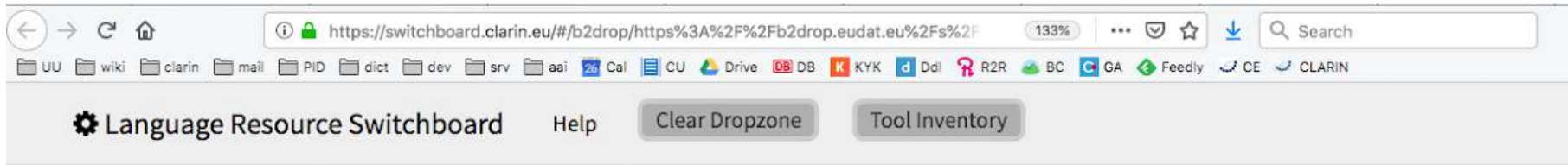
Allow editing

Password protect

Set expiration date

Więcej: <https://www.clarin.eu/showcase/eosc-portal-demonstration>

CLARIN Switchboard and EOSC



Resource transferal from B2DROP. Please check the information below, then press "Show Tools"

Input Analysis

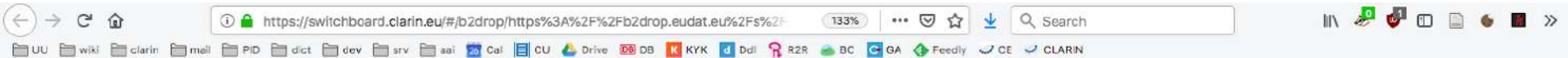
resource	mimetype	language
name: :download?input=https::b2drop.eudat.eu:s:dFexm9tS7TJdoZE:download size: 5427273 bytes	application/zip	Polish
	<input type="button" value="▼"/>	<input type="button" value="▼"/>
		<input type="button" value="Show Tools"/>

About
v1.3.0-pro/docker (Oct 8, 2018)

For Developers
Service provided by CLARIN

<https://switchboard.clarin.eu/>

EOSC → Switchboard → WebSty (CLARIN-PL)



Tools

Only Tools

Both Tools & Web Services

Only Web Services

Sort by Task

Order Alphabetically

Stylometry

WEBSTY



Similarity and clustering of texts in Polish. The tools used include: Morfeusz 2 with SGJP dictionary (for morphological analysis), wcrft2 (for tagging), Liner2 (for named entities recognition), Fextor (for extraction of features from texts); Cluto (for clustering), result visualisation: D3.js, D3-tip.

http://ws.clarin-pl.eu

no

application/octet-stream

Click to start tool

Wrocław, Poland

tomasz.walkowiak@pwr.edu.pl

<https://switchboard.clarin.eu/>

WebSty: processing the corpus

<http://ws.clarin-pl.eu/websty.shtml>

- Default settings for different types of research tasks

Opcje podstawowe	Ustawienia wstępne
LICZBA GRUP ⓘ <input type="text" value="2"/>	METODY ANALIZY ⓘ <input type="text" value="Analiza autorstwa"/> ▼
<input checked="" type="checkbox"/> PODZIAŁ PLIKÓW WEJŚCIOWYCH ⓘ <input type="text" value="20000"/> ⬆️⬆️	<input type="checkbox"/> PONOWNE WYKORZYSTANIE CECH <input type="text" value="/resources/fextor/5autorow/kaa"/>
	ŹRÓDŁO WEKTORA CECH <input type="text" value="ID z ostatniej analizy"/> ▼

Ustawienia wstępne	
METODY ANALIZY ⓘ	<input checked="" type="checkbox"/> Analiza autorstwa <input type="checkbox"/> Analiza stylu gramatycznego <input type="checkbox"/> Grupy podobieństwa treści <input type="checkbox"/> Klasyczna analiza autorstwa
<input type="checkbox"/> PONOWNE WYKORZYSTANIE CECH	
ŹRÓDŁO WEKTORA CECH	<input type="text" value="ID z ostatniej analizy"/> ▼

WebSty: processing the corpus

<http://ws.clarin-pl.eu/websty.shtml>

Wybór cech ^

GRAMATYCZNE I LEKSYKALNE SŁOWNIKOWE MODELOWANIE TEMATYCZNE WEKTORY DYSTRYBUCYJNE

Elementy

LEMATY ? ▼ FORMY WYRAZOWE ? ▼

Interpunkcja

DOWOLNE ZNAKI POSZCZEGÓLNE ZNAKI Z LISTY ?

Części mowy ?

CZASOWNIKI PRZYMIOTNIKI PRZYIMIKI

RZECZOWNIKI PRZYŚLÓWKI

Pozostałe klasy gramatyczne ?

RZECZOWNIKI POSPOLITE (SUBST W NKJP) FORMY WINIEN PSEUDOIMIESŁOWY

FORMY DEPRECJATYWNE PREDYKATY ROZKAŹNIKI

WebSty: visualisation of the results



INTERAKTYWNY DENDROGRAM



MAPA CIEPŁA



SKALOWANIE WIELOWYMIAROWE



SKALOWANIE WIELOWYMIAROWE Z
WIZUALIZACJĄ 3D



WYKRES RADAROWY



WYKRES KOŁOWY



PLIK XSLX



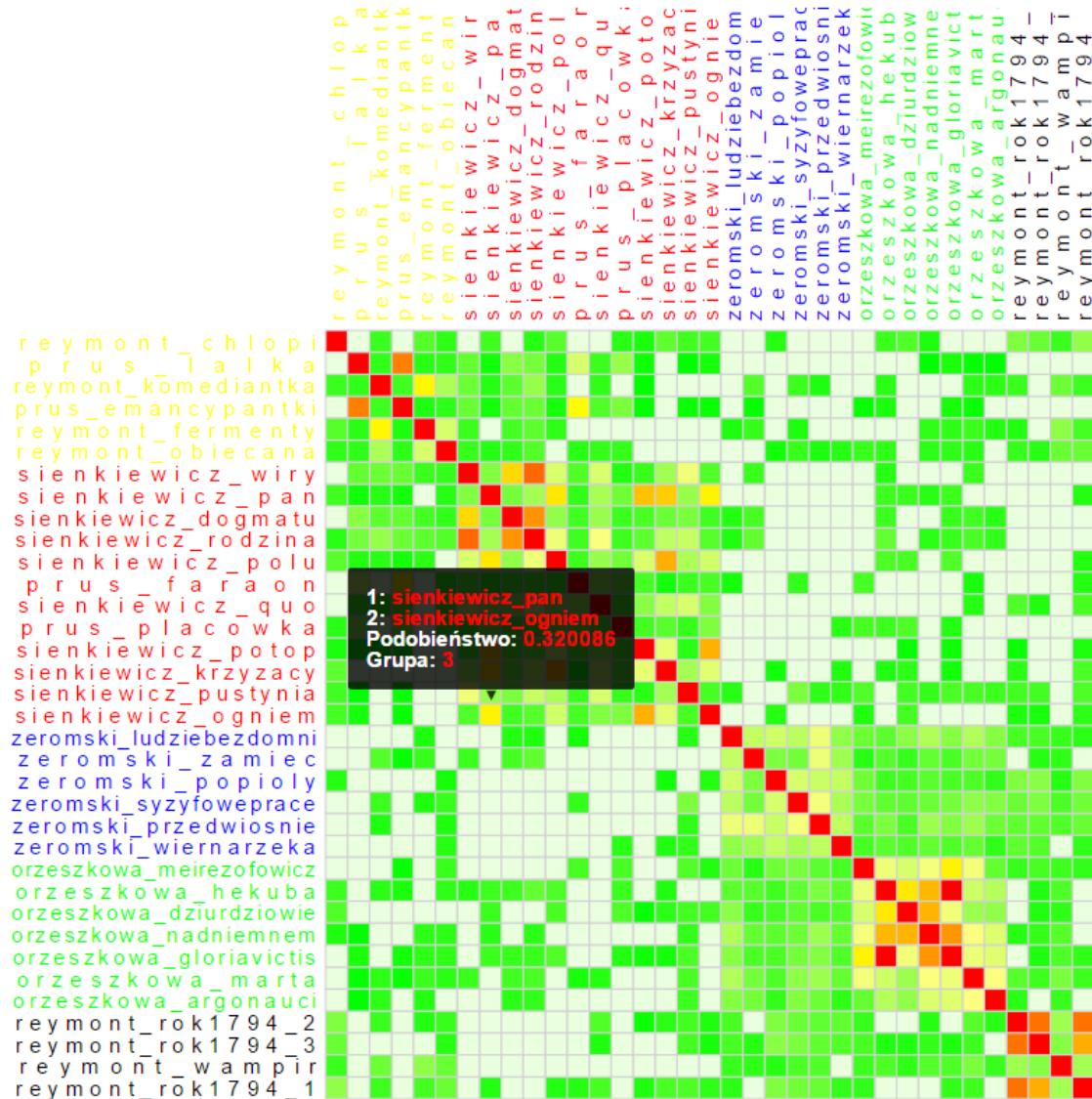
ISTOTNOŚĆ CECH



REZULTATY



Results presentation – ‘the warmth map’



Multidimensional scaling

METODA SKALOWANIA

TSNE

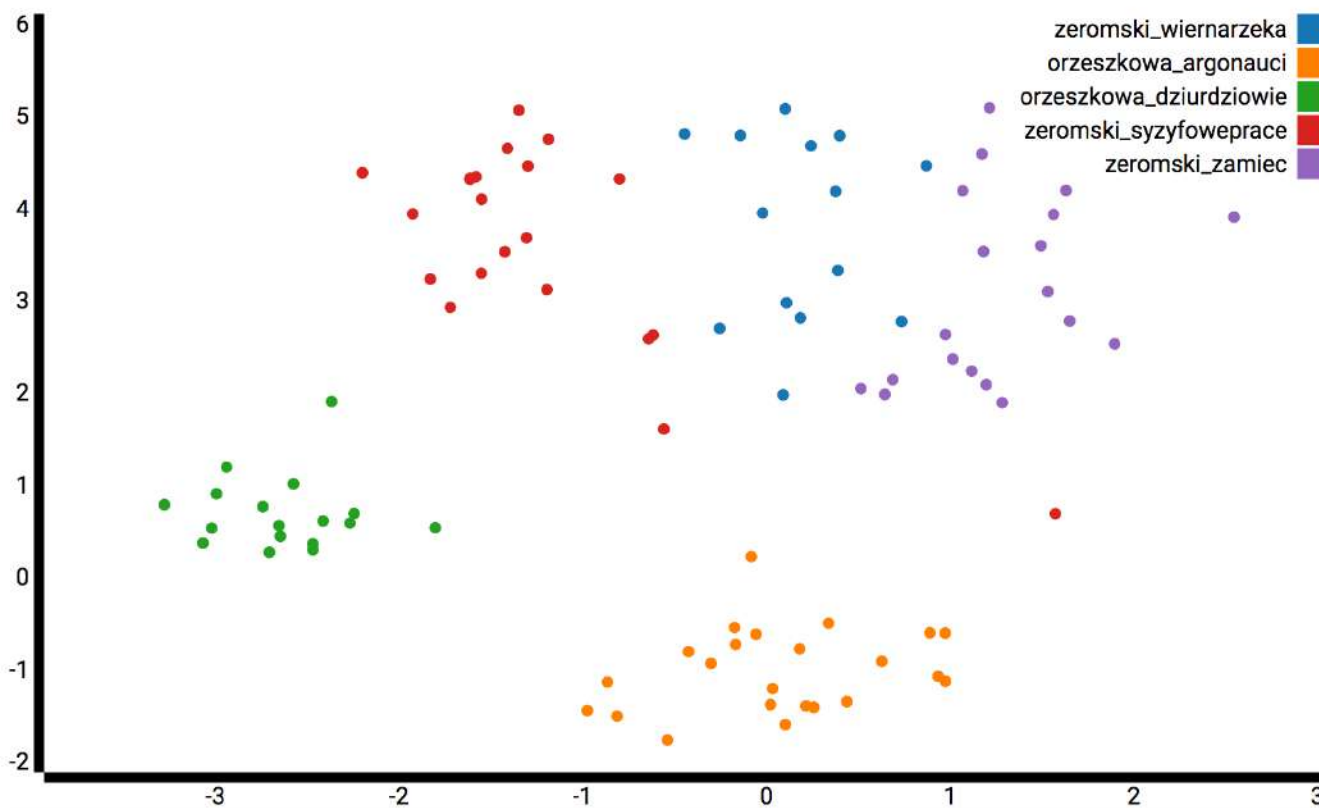


PERPLEXITY

40



Przelicz



ETYKIETY

PODZIAŁ NA GRUPY

ostatni poziom



ETYKIETY GRUPY



Press agency news

- Vector model, cosine, t-SNE



CLARIN Poland: CLARIN-PL

(since 2006, in ERIC since 2012)



CLARIN-PL language Technology Centre
(<http://clarin-pl.eu>)

language data repository



CLARIN Cloud – private data cloud for researchers
(<https://nextcloud.clarin-pl.eu/>)

language resources for Polish and other languages
services and applications for text and speech analysis
(<https://ws.clarin-pl.eu>)

PolLinguaTec – Knowledge Centre for Language Technology
for the Polish Language

<http://kcentre.clarin-pl.eu/>



CLARIN-PL – consortium for open Polish LT

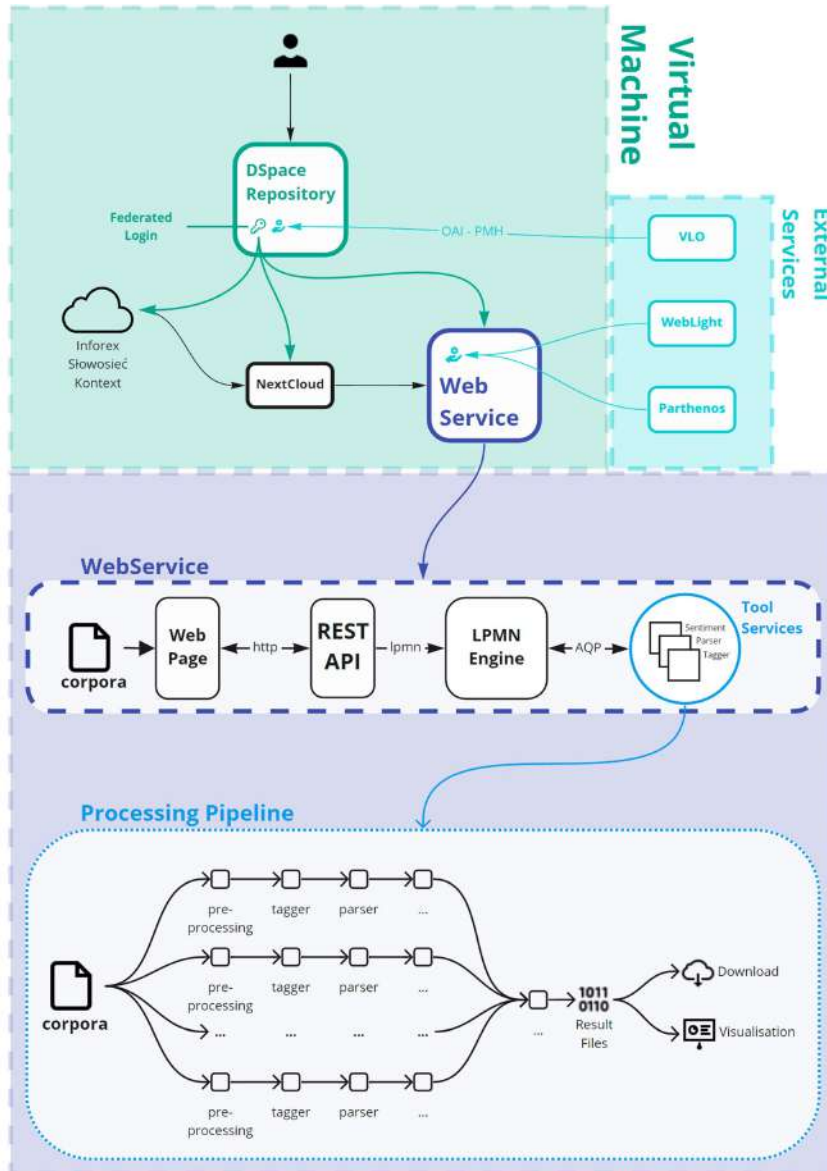
- Developers (consortium):
 - Department of Artificial Intelligence, Wrocław University of Science and Technology (leader)
 - Institute of Computer Science PAS
 - Institute of Slavistics, PAS
 - Polsko-Japanese Academy of Information Technology
 - University of Łódź
 - University of Wrocław
- Beneficiaries:
 - **All** research units and **Researchers** in Poland, especially from the area of Social Sciences and Humanities
 - also **Artificial Intelligence** in broad sence
(due to the CLARIN-PL-Biz project: www.clarin.biz)

CLARIN-PL – open support for researchers and beyond

- **Open science (FAIR):**
 - language resources and data, software and applications
 - knowledge and direct support
 - **no fees** – sponsored by **Ministry of Education and Science** (than you!)
- Language resources:
 - corpora: of Polish and multilingual, richly annotated
 - Lexical data bases: morphological, grammatical, semantic – among the largest in the world, e.g. plWordNet connected to Linked Open Data
- Basic language tools
 - morphological analysis, grammatical, Information Extraction.
- Research applications
 - development and analysis of corpora, statistical analysis, stylometry, sentiment and emotions, semantic analysis
- Direct support for researchers:
 - teams, projects and individual researchers and students
 - from an idea, via problem definition, tasks, project application till support in conducting research

CLARIN-PL Language Technology Centre

ws.clarin-pl.eu

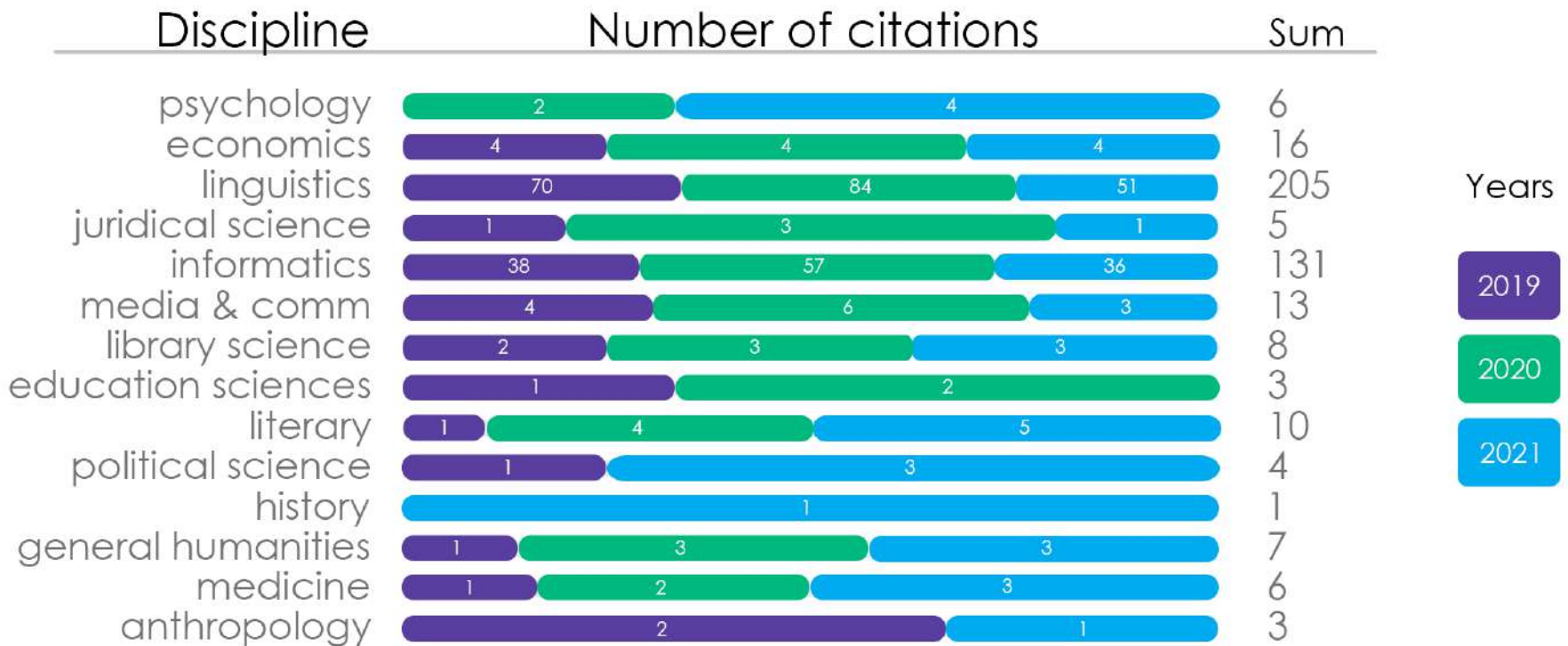


- Robust
- Efficient
- Flexible orchestration of NLP processing pipelines
- Parallel processing
- Dynamic scaling
- Millions requests served per year

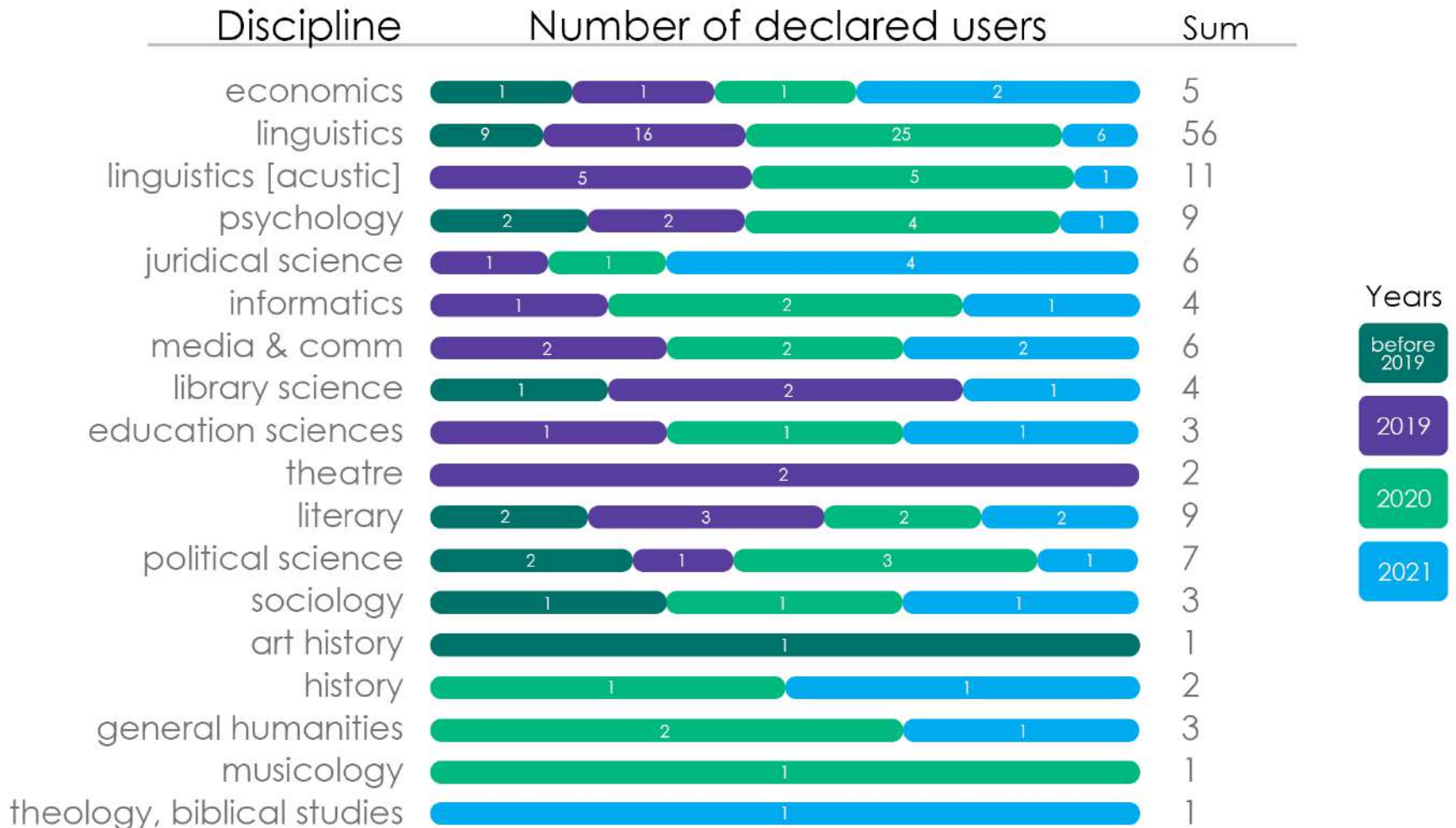
CLARIN-PL Users (VI 2018 – VI 2021)

- Known users of CLARIN-PL (resources, tools and applications):
 - **199 teams and individual researchers, in total**
 - activities:
 - from using resources and tools, till CLARIN-PL team involvement (**free!**) in works on expansion of resources, adaptation of tools and construction of new versions of applications for the users
- Spontaneous users: 442
 - known from citations and different mentions in the web
- Users from education: 94
- Use statistics:
 - 6 871 828 processing tasks only on ws.clarin-pl.eu
 - 25 195 294 processed documents,
 - of **the total volume: 405 736 [QV = QuoVadis]**
 - 1 [QV] = amount of text in „Quo Vadis” of Henryk Sienkiewicz

CLARIN-PL Users – citations



CLARIN-PL Users



EOSC SSH Cluster powered by CLARIN

- National consortia and national LT research infrastructures integrated into CLARIN ERIC
 - VLO – meta-repository based on common metadata standard
 - Federated Content Search - corpora
 - Language Switchboard – tools and applications
- CLARIN ERIC services linked to various EOSC initiatives
 - ready to use solutions
 - already tested in prototypes
- CLARIN ERIC cooperation with other SSH RI on different levels
- A blueprint for bottom-up construction of EOSC via already implemented collaboration
 - richness and strength of CLARIN contributing to Pan-European cooperation

Thank you for your attention!

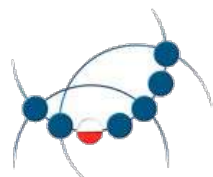
Learn more about CLARIN at:

www.clarin.eu
clarin-pl.eu

or

clarin@clarin.eu
clarin-pl@pwr.edu.pl
maciej.piasecki@pwr.edu.pl

CLARIN-PL
Common Language Resources and Technology Infrastructure



<http://clarin-pl.eu/>
<http://clarin.biz>



<http://clarin.eu/>