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

Maciej Piasecki

CLARIN-PL

Wrocław University of Science and Technology

CLARIN ERIC

CLARIN

CLARIN-PL
Common Language Resources and Technology Infrastructure

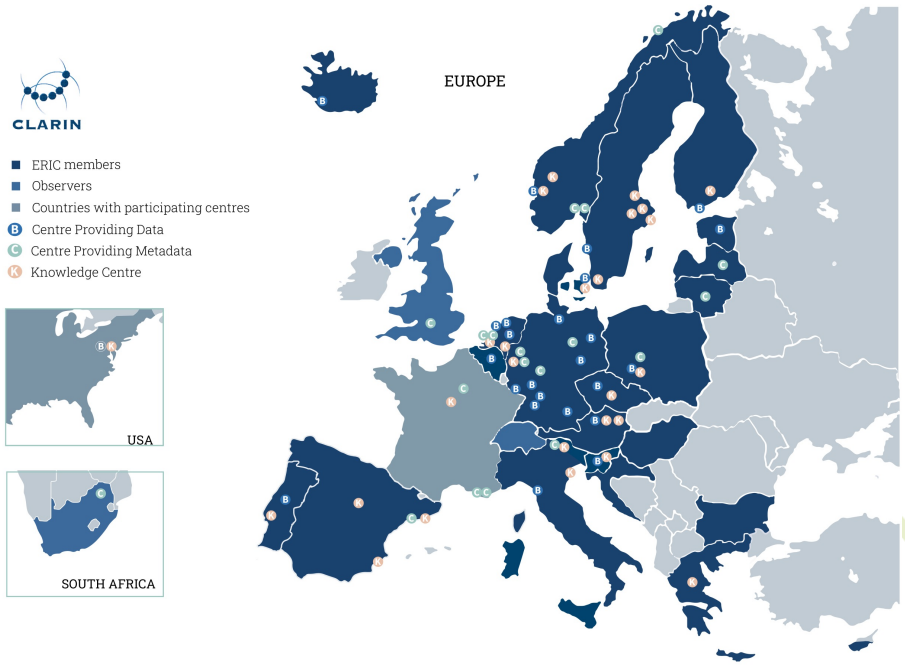
Wrocław University of Science and Technology

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**

- 24 member countries
- 2 observers
- 1 linked party



A distributed network of 48 centres

21 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 eosc catalogue

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

- Multidisciplinary collaboration in EOSC Future (H2020) EOSC Future 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

SSHOpenCluster
Social Sciences & Humanities Resources

Interconnecting existing and new infrastructures

CLARIN-PL Consortium for Open Polish LT

- Developers (consortium) – institutions supporting Open Science in Language Technology (at 2006):
  - Department of Artificial Intelligence, Wrocław University of Science and Technology (leader)
  - Institute of Computer Science PAS
  - Institute of Slavistics, PAS
  - Polish-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 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 (old: https://ws.clarin-pl.eu **new:** https://services.clarin-pl.eu)

PolLinguaTec – Knowledge Centre for Language Technology for the Polish Language

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

CLARIN-PL – open support for researchers and beyond

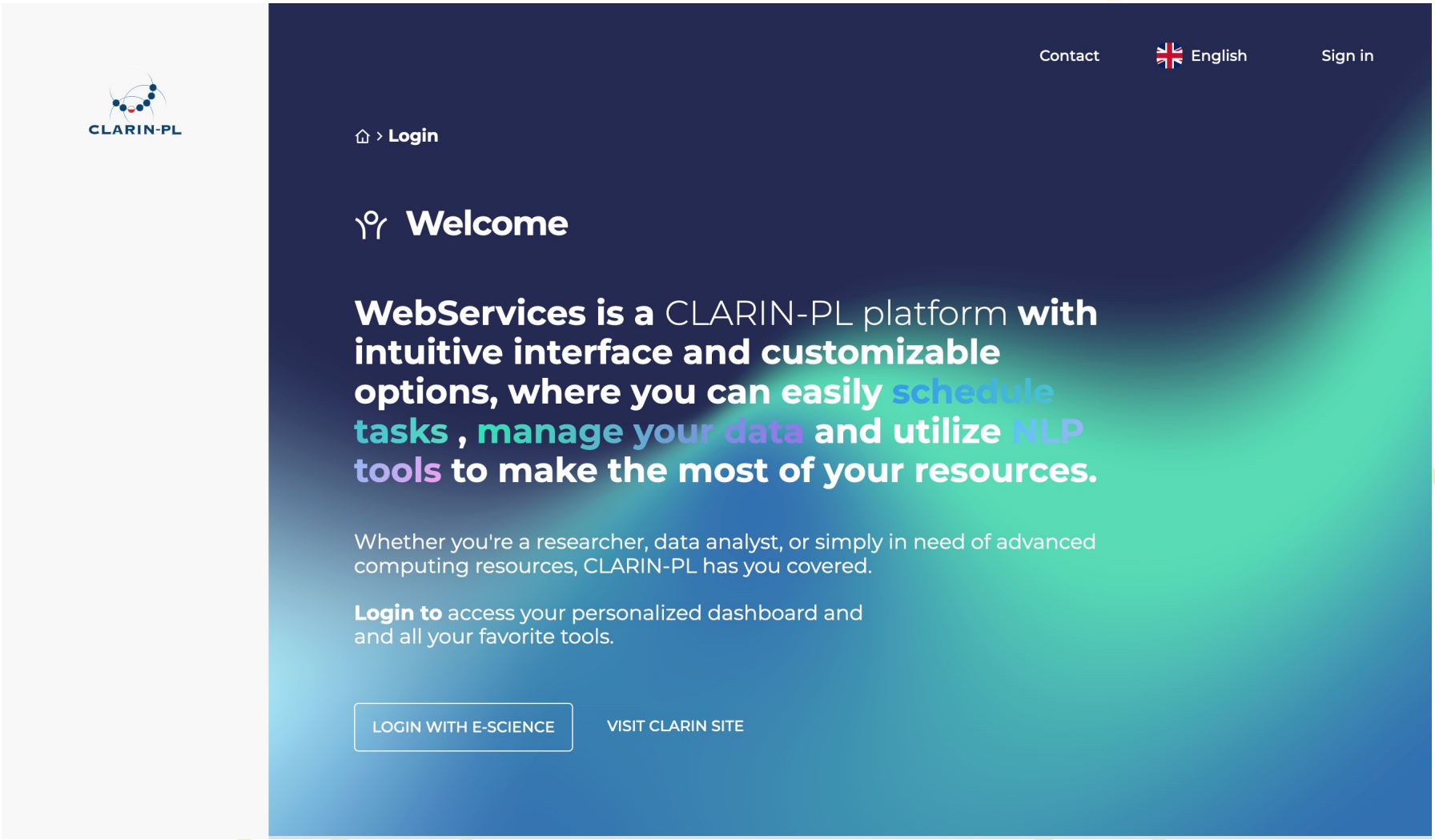
- **Open science** (FAIR):
  - language resources and data, software and applications
  - knowledge and direct support
  - **no fees** – sponsored by **Ministry of Education and Science** (than you!)
- Language resources:
  - corpora: of Polish and multilingual, richly annotated
  - lexical data bases: 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 for Research & Development

- CLARIN-PL-Biz (2020–2023) financed from The Smart Growth Operational Programme (POIR 4.2) - a large, strategic research infrastructure, :
  - 5 research units (CLARIN-PL) and 28 Business Partners
  - 31.3 mln. € : 24.14 mln. € of support and 4.55 mln. € in-kind contribution from Business Partners
- Results
  - 60% open LT: fundamental resources an tools
  - 40% paid access: specialised services built according to requirements collected from business partners
  - new CLARIN-PL technological centre – **an HPC for NLP**
  - language datasets for Machine Learning
  - language models, including generative models
  - language tools, applications and services for science and business

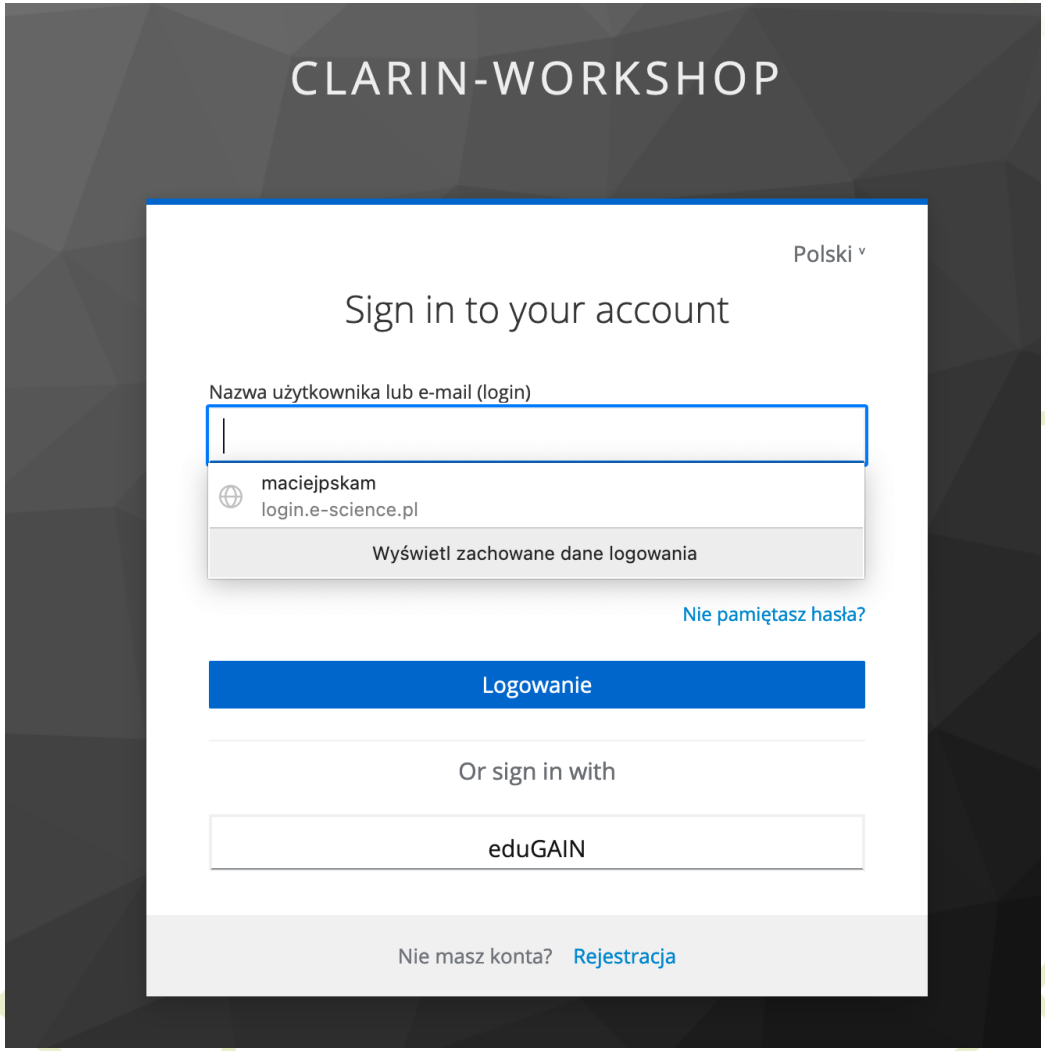
CLARIN-PL (Biz) via Federated Login

https://services.clarin-pl.eu



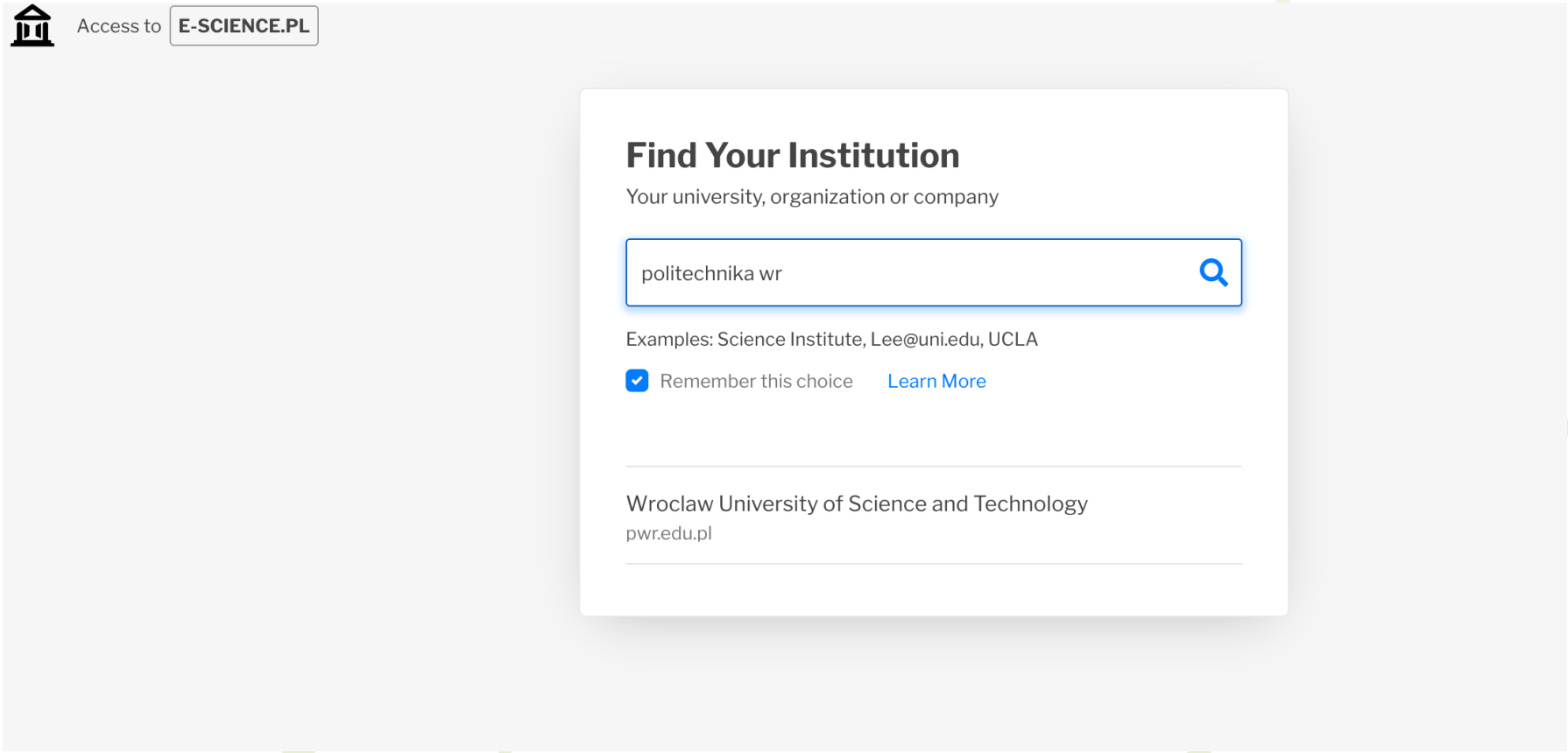
CLARIN-PL (Biz) via Federated Login

https://services.clarin-pl.eu



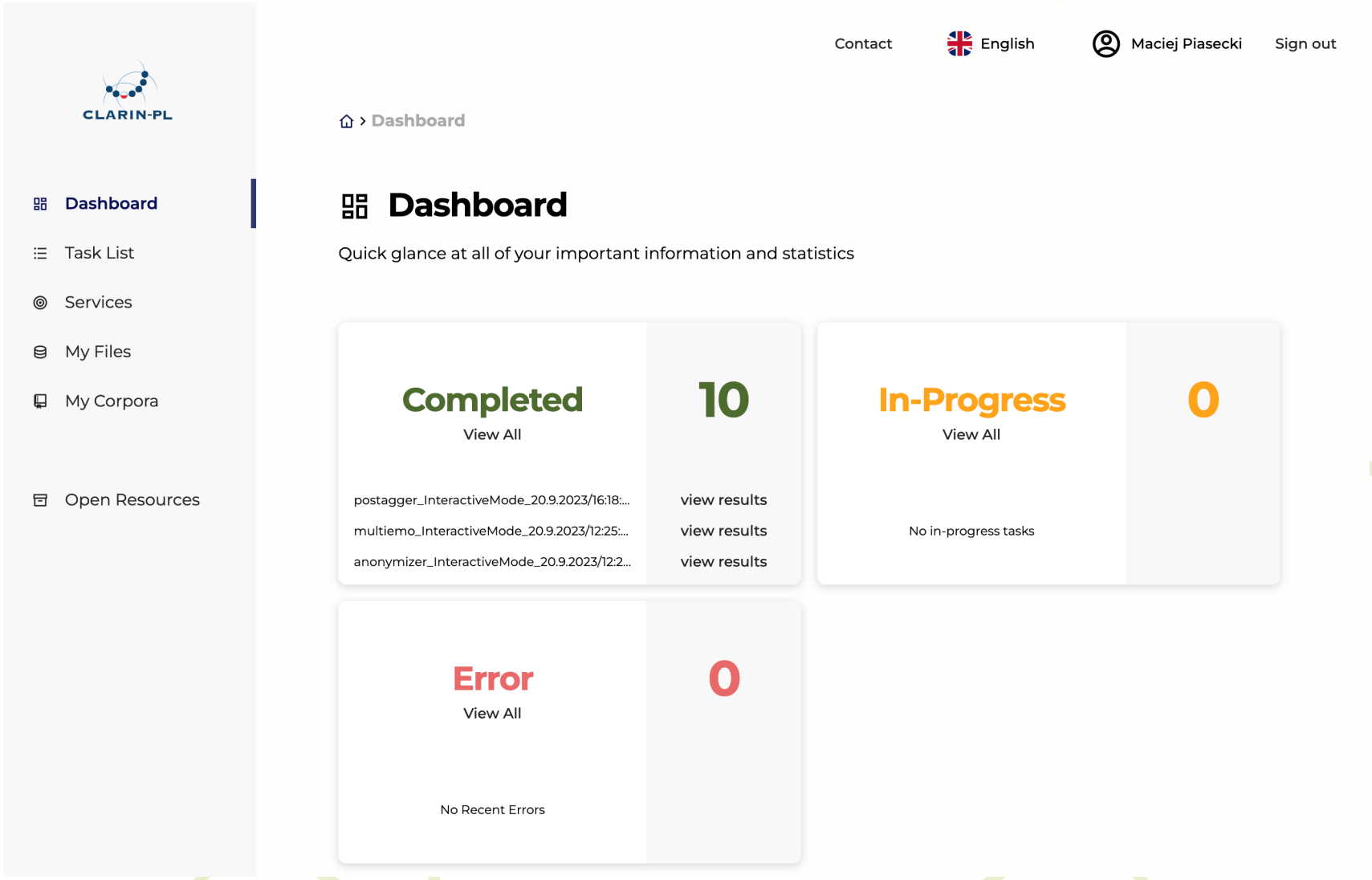
CLARIN-PL (Biz) via Federated Login

https://services.clarin-pl.eu



CLARIN-PL (Biz) via Federated Login

https://services.clarin-pl.eu



CLARIN-PL (Biz) via Federated Login

https://services.clarin-pl.eu

Services › Multiemo › **Interactive**

Interactive Mode

MultiEmo

Input

Paste text to be processed and adjust parameters of your task

Text | Parameters

Zamkniesz się już czy kopnąć Cię w dupę? Oczywiście nie dlatego, że jesteś grubą świnią.

Start ▶

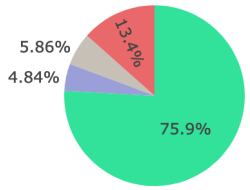
Output

Select a result type to preview it

json | distribution-list

Negative

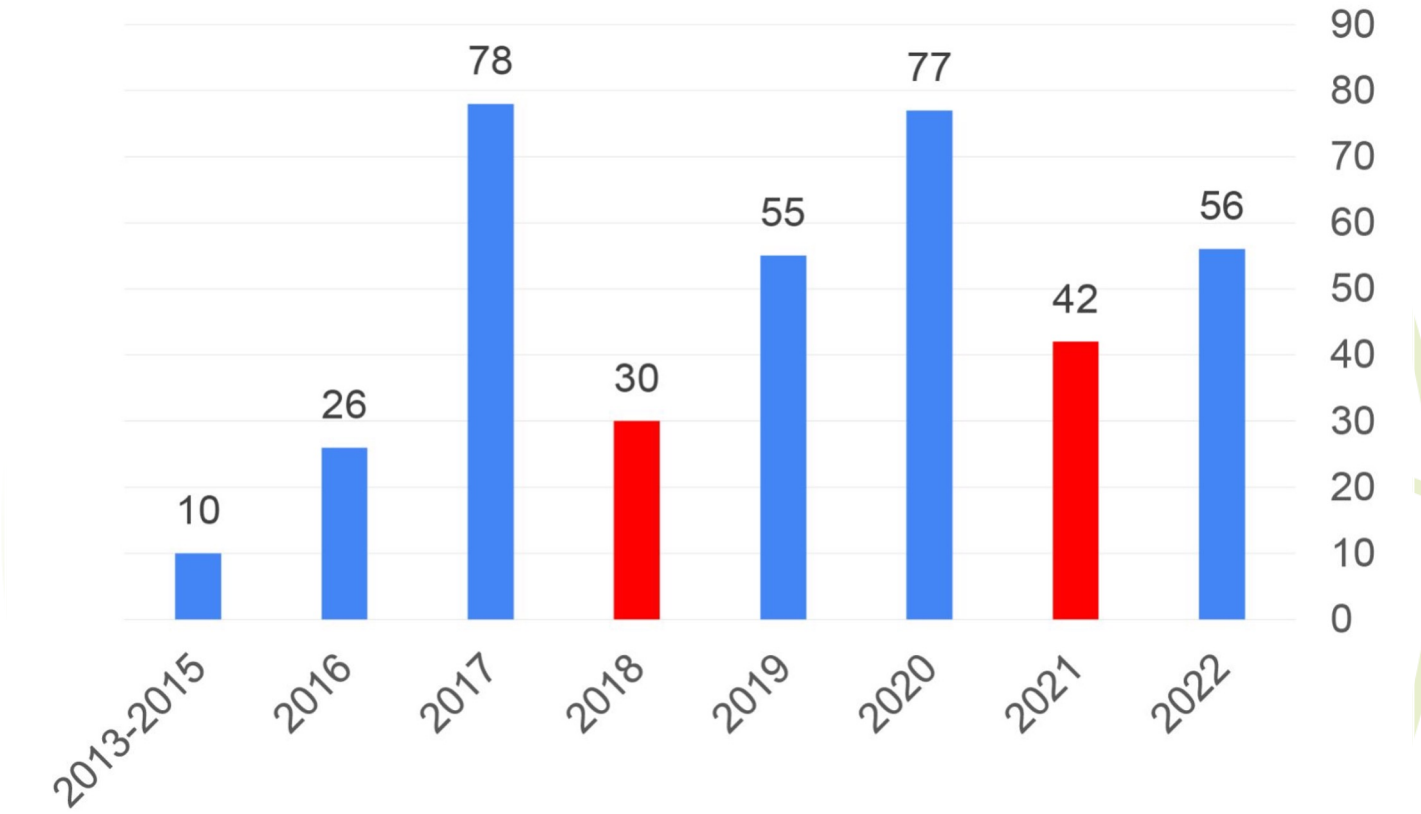
Zamkniesz się już czy kopnąć Cię w dupę? Oczywiście nie dlatego, że jesteś grubą świnią.



CLARIN-PL for Research & Development

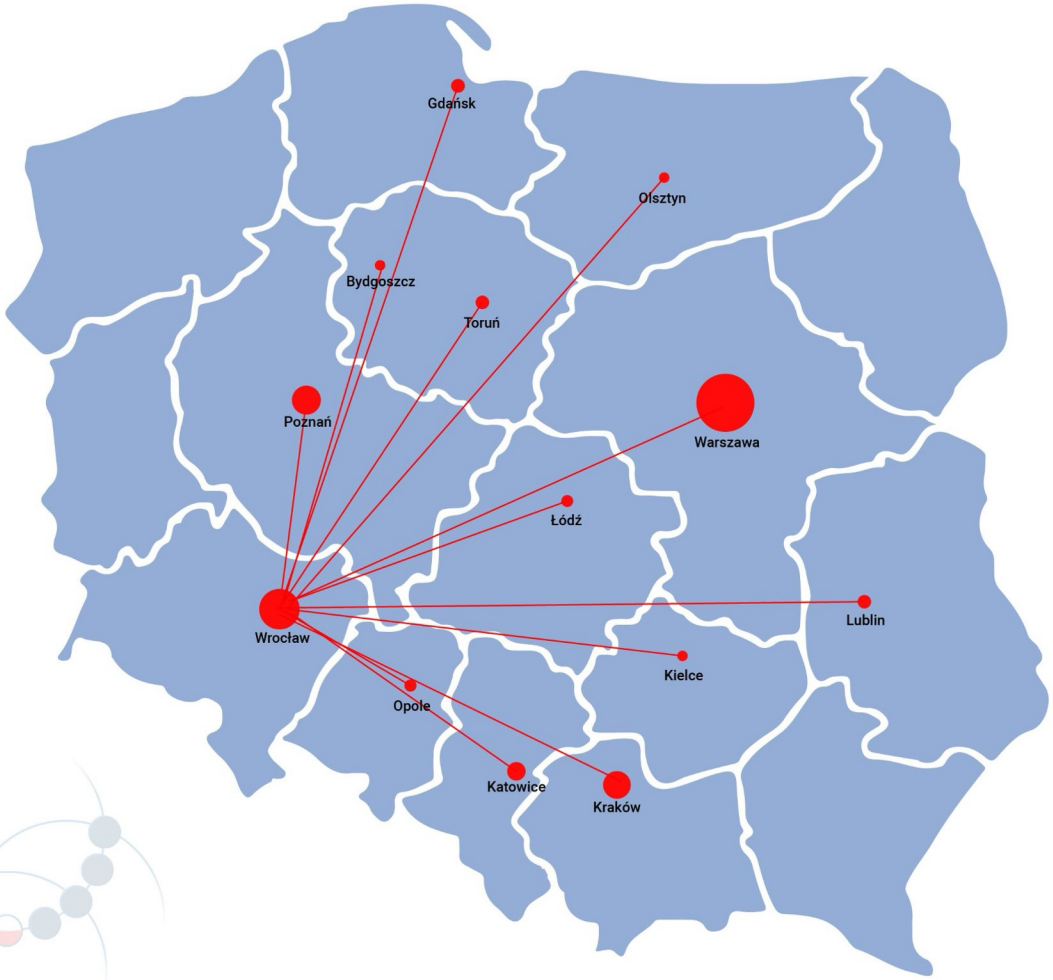
- Text correction, normalisation and intelligent anonymisation
- Lepiszcze (http://lepiszcze.ml/) – a comprehensive benchmark dataset and test board (*leaderboard*)
- Grammatical analysis and parsing
- Text-to-knowledge resources mapping
- Personalised analysis of sentiment and emotions
- Architecture of trusted dialogue systems based on LLMs
- Information Extraction and text data streams analysis
- Advanced Question Answering and Natural Language Inference
- Stylometry and semantic classification of texts

CLARIN-PL Users: Known Users



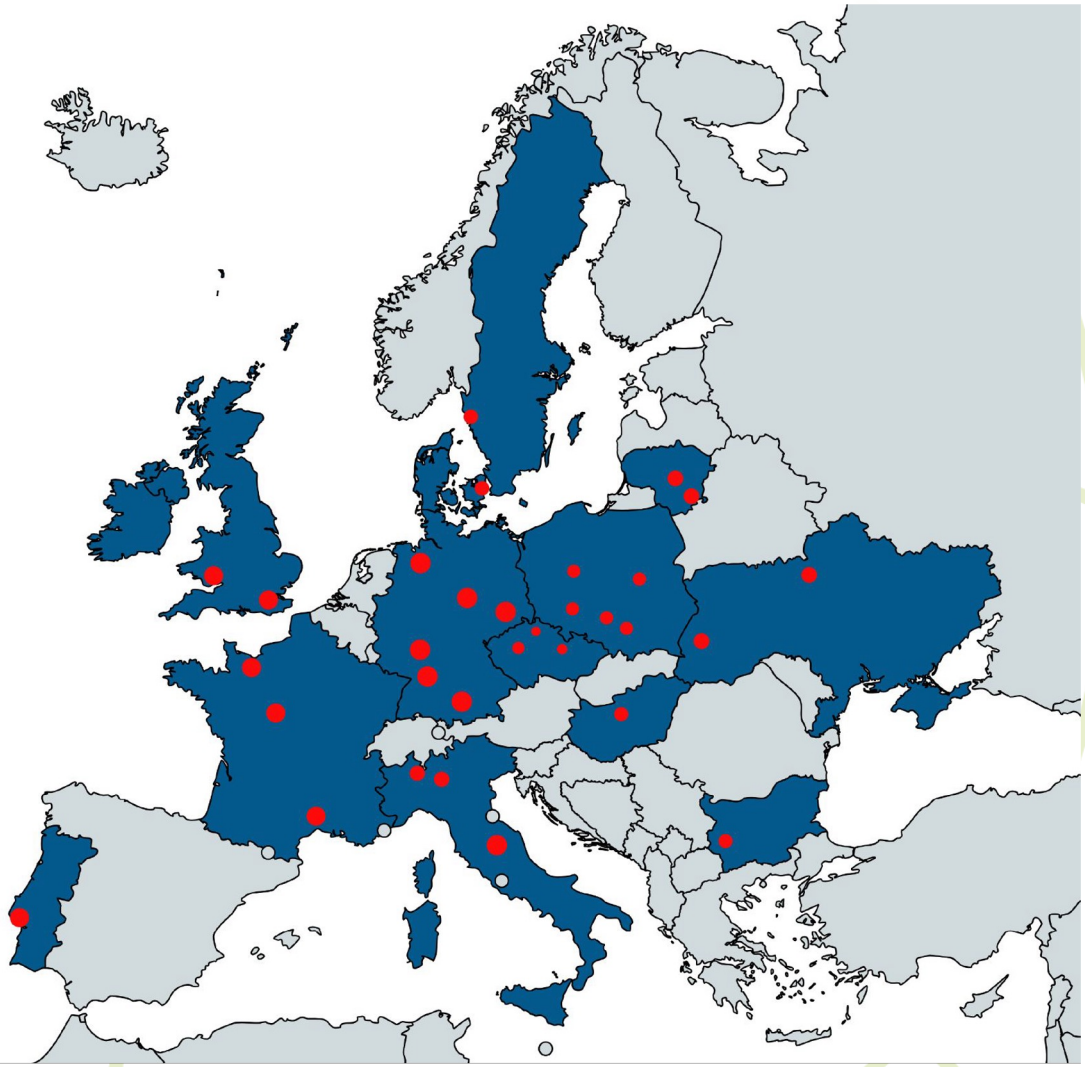
- Users in direct contact: research teams and individuals
- Adaptation of tools or applications for them

CLARIN-PL Users: Known Users

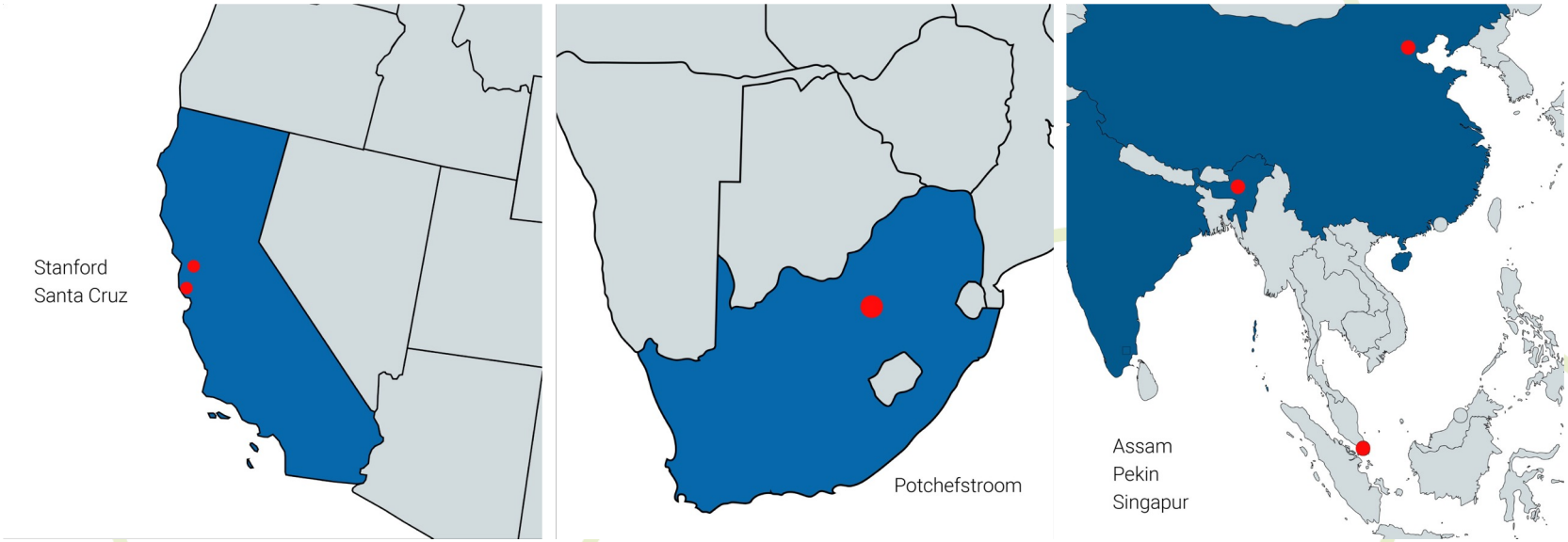


CLARIN-PL Users: Known Users

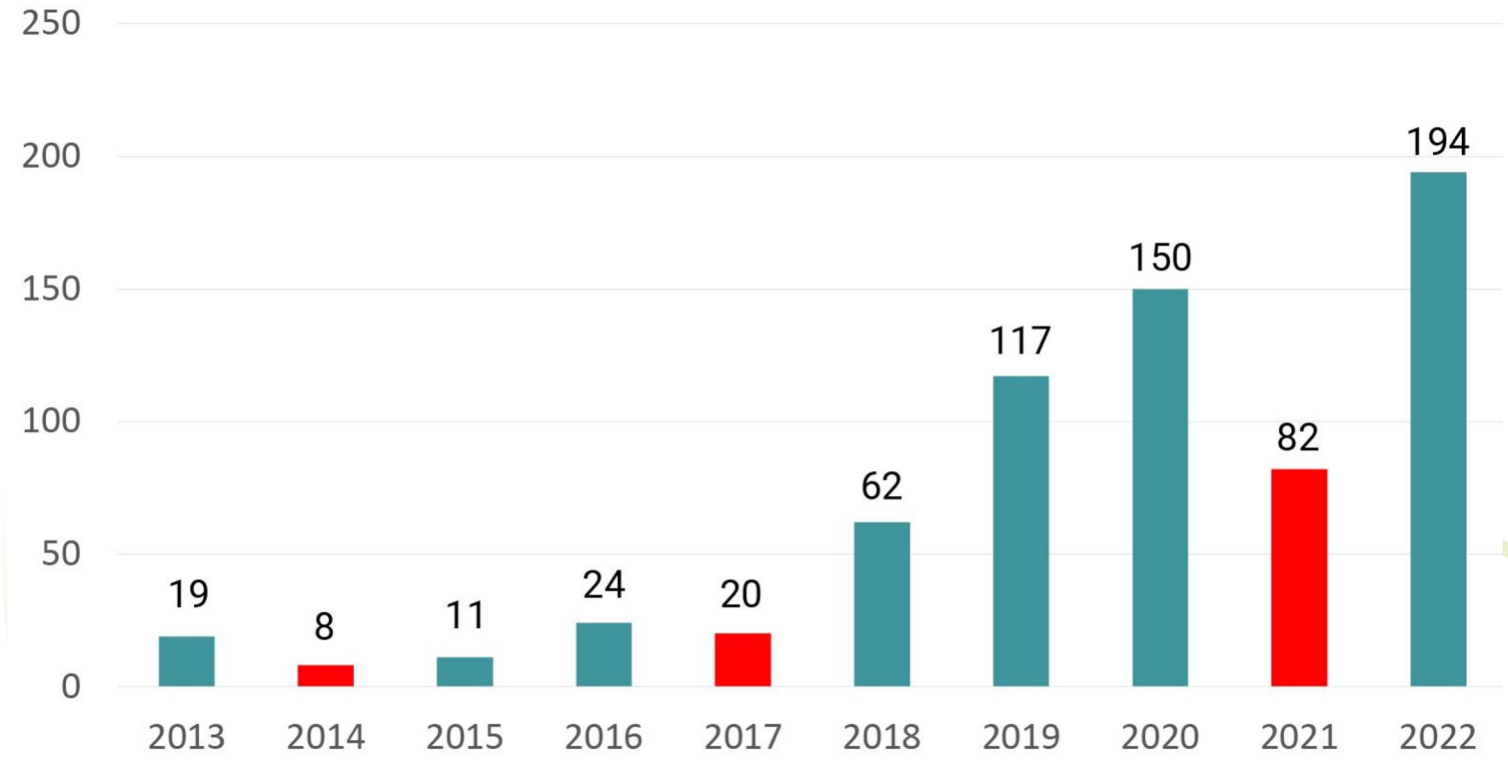
Bułgaria
Czechy
Dania
Francja
Irlandia
Litwa
Niemcy
Polska
Portugalia
Szwecja
Ukraina
Wielka Brytania
Węgry
Włochy



CLARIN-PL Users: Known Users

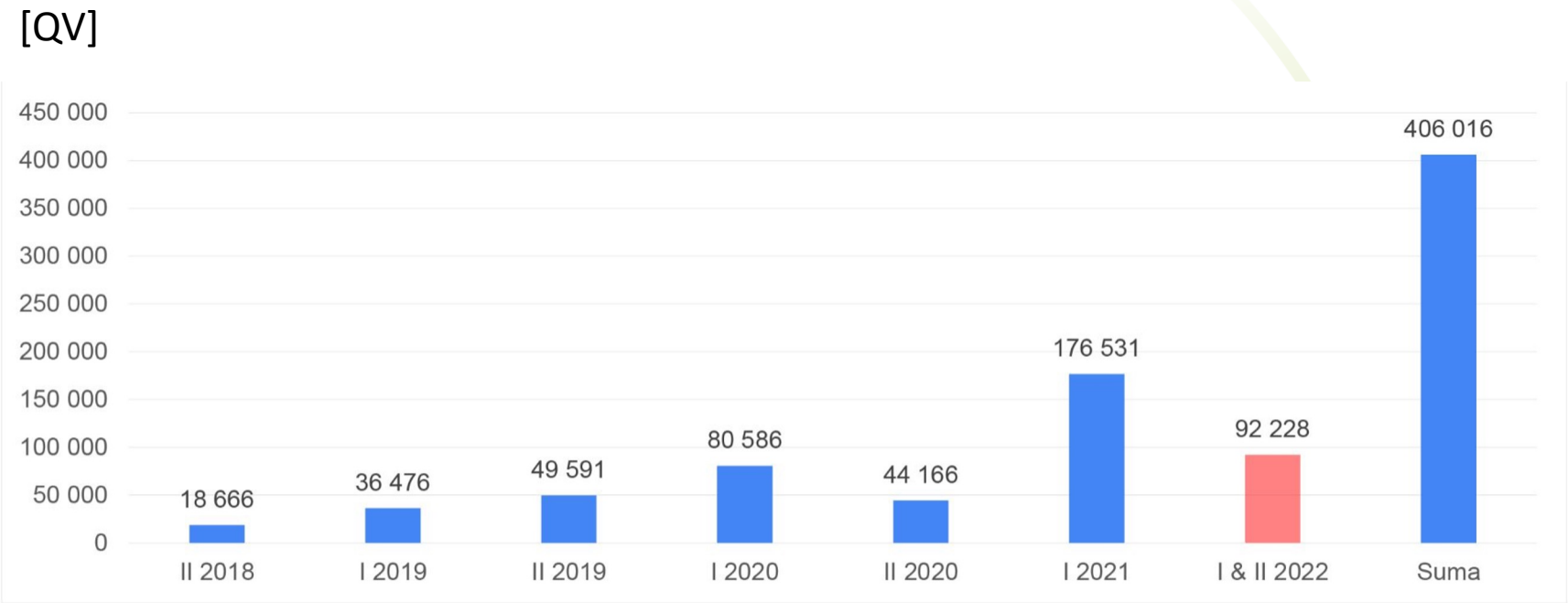


CLARIN-PL Users: Spontaneous Users



- 687 in total
- Identified on the basis of citations or mentions

CLARIN-PL Users: Language Data Processed



- Total volume of language data measured in QuoVadises [QV]
- 1 [QV] = amount of text in „Quo Vadis” of Henryk Sienkiewicz

CLARIN-PL Users: Disciplines from Citations

DYSCYPLINA	2019	2020	2021	2022	ŁĄCZNIE
lingwistyka	70	84	51	51	256
informatyka	38	57	36	16	147
ekonomia i zarządzanie	4	7	5	7	23
badania medioznawcze i komunikologiczne	4	6	3	7	20
ogólnohumanistyczne	1	3	3	8	15
literaturoznawstwo	1	4	5	3	13
bibliotekoznawstwo i informacja naukowa	2	3	3	3	11
psychologia	0	2	4	2	8
medycyna	1	2	3	2	8
politologia i stosunki międzynarodowe	1	0	3	3	7
prawo	1	3	1	1	6
etnologia	2	0	1	3	6
nauki o edukacji	1	2	0	2	5
historia sztuki	0	1	1	1	3
socjologia	0	0	0	2	2
historia / archiwistyka	0	0	1	1	2
ŁĄCZNIE	126	174	120	112	532

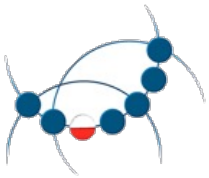
CLARIN – Lessons Learned

- A blueprint for bottom-up construction of EOSC via already implemented collaboration:
  - EOSC SSH Cluster powered by CLARIN
  - VLO – meta-repository based on common metadata standard
  - Language Switchboard – tools and applications
- Open solutions with potential applications to other RIs
- Specific and diversified needs of SS&H researchers, strongly inspired by recent developments in the generative AI
- CLARIN-PL model of continuous cooperative development and active support for research users
- Challenges:
  - sustainability of the active support for users and EOSC
  - continuous maintenance, evolving technology and research needs

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/