

Otwarcie od zawsze - otwarta nauka a infrastruktura CLARIN-PL

Jan Wieczorek

CLARIN-PL

Wrocław University of Science and Technology

CLARIN ERIC



CLARIN-PL
Common Language Resources and Technology Infrastructure



Wrocław University
of Science and Technology

CLARIN ERIC - członkowie i centra

Infrastruktura naukowa CLARIN (= Common Language Resources and Technology Infrastructure)

Konsorcjum typu ERIC (od 2012) - **Social Sciences and Humanities Cluster**

- 25 krajów członkowskich
- 3 obserwatorów
- 1 podmiot współuczestniczący

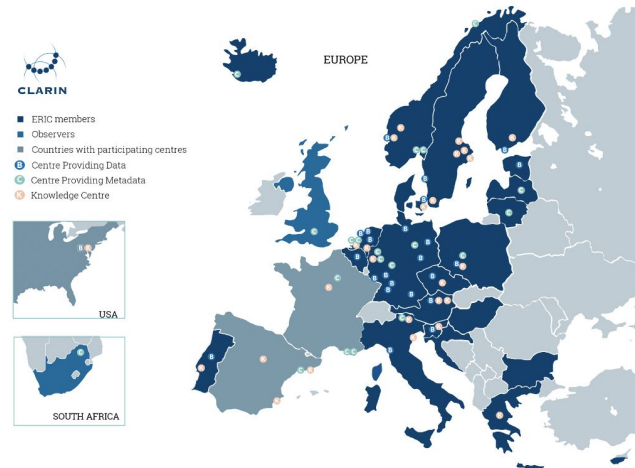
Rozproszona sieć złożona z 60 centrów

25 z nich posiada certyfikat CTS,

Silny nacisk na standardy FAIR i interoperacyjność

- logowanie federacyjne: 
- scentralizowane pobieranie metadanych 
- zintegrowane usługi: 
- dane językowe - pisane, mówione, wideo lub w formie multimodalnej
- zaawansowane narzędzia - dla odkrywania, eksplorowania, wykorzystywania, anotowania, analizowania i łączenia danych - gdziekolwiek są zlokalizowane

Członek EOSC-Association  **eOSC** (od 2020)



CLARIN w Polsce: CLARIN-PL

(od 2006, w ERIC od 2012)

- CLARIN-PL Language Technology Centre
(<http://clarin-pl.eu>)
 - repozytorium danych językowych
 - CLARIN Cloud – prywatna chmura danych dla naukowców (<https://nextcloud.clarin-pl.eu/>)
 - zasoby językowe dla języka polskiego i innych
 - narzędzia i aplikacje do analizy tekstu i mowy (<https://ws.clarin-pl.eu>)
- PolLinguaTec – Knowledge Centre for Language Technology for the Polish Language
 - <http://kcentre.clarin-pl.eu/>

CLARIN-PL – konsorcjum otwartej LT dla polszczyzny

- Wytwórcy (konsorcjanci):
 - Katedra Sztucznej Inteligencji, Wrocław University of Science and Technology (lider),
 - Zespół Inżynierii Lingwistycznej, Instytut Podstaw Informatyki PAN,
 - Instytut Slawistyki, PAN,
 - Polsko-Japońska Akademia Technik Komputerowych
 - Instytut Anglistyki, Uniwersytet Łódzki,
 - Uniwersytet Wrocławski.
- Beneficjenci:
 - **Wszystkie jednostki naukowe i naukowcy w Polsce** - szczególnie z obszaru nauk społecznych i humanistycznych
 - również **Sztuczna Inteligencja** w szerokim sensie (dzięki projektowi CLARIN-PL-Biz: www.clarin.biz)

CLARIN-PL: priorytety w kontekście ON

- tworzenie nowych otwartych zasobów językowych
 - zasoby są ważniejsze od narzędzi i aplikacji, bo od ich jakości zależy później jakość innych elementów infrastruktury; zasoby powstają zazwyczaj dłużej niż narzędzia i aplikacje, wymagają większych nakładów pracy;
- wsparcie techniczne i know how w tworzeniu nowych zasobów otwartych przez użytkowników
 - corocznie pomagamy w tworzeniu kilku nowych zasobów otwartych - od etapu koncepcji po publikację zasobu oraz jego promowanie w społeczności akademickiej
- utrzymywanie dostępu do zasobów własnych i partnerów zewnętrznych
 - zasoby są deponowane oraz rozpowszechniane za pomocą infrastruktury CLARIN, służy do tego repozytorium oraz serwis VLO i Federated Content Search
- staramy się publikować własne teksty w otwartym dostępie (nie zawsze możliwe)
- prowadzimy działalność szkoleniową, która również ma promować publikowanie zasobów na zasadach FAIR

CLARIN-PL: otwarte zasoby

- Korpus Politechniki Wrocławskiej
- Chronopress (korpus dawnych tekstów prasowych)
- esPresso (powstaje)
- Korpus Dyskursu Parlamentarnego
- Wielojęzyczny korpus równoległy (PL-BUL-LT-UKR-RUS)
- Paralela - Korpus tłumaczeń PL-ENG/ENG-PL
- Serwis Frazeo (korpus monitorujący polski Internet)
- Korpus Czterech Wieszczów (powstaje, wraz z IBL)
- Spokes - Korpus mowy spontanicznej
- Polish Spatial Texts Corpus

CLARIN-PL: otwarte zasoby leksykalne

- Słownosieć - plWordNet
- Słownosieć EN / Open English WordNet, Open MultiLingual WordNet
- Walenty - słownik walencyjny języka polskiego
- mapowanie na różne ontologie (SUMO, ontologie dziedzinowe, JHP Biblioteki Narodowej itd.)
- Słownosieć wzbogacona anotacją emocjonalną

CLARIN-PL: otwarte modele językowe

- Modele publikowane na Huggingface w otwartym dostępie:
 - rozpoznawanie słów kluczowych w tekstach
 - ulepszone modele do ujednoznaczniania znaczeń słów (WSD)
 - rozpoznawanie bytów nazwanych (NER)

CLARIN-PL: wsparcie powstawania zasobów

- Korpus Listów Pożegnalnych Samobójców
- Korpus Języka Mówionego Wrocławia
- POLITORATOR - Korpus Polskich Przemówień Politycznych 1918 - 2018
- 19/20 MetaPNC (Polish Novel Corpus)
- Korpus Dawnego Dramatu Polskiego (tytuł roboczy)

Ale też na przykład:

- Street name changes in Poznań, Słubice and Zbąszyń, Poland 1916-2018
- zasoby dla języka assamskiego
- Metaphors and arguments annotations in Polish political – pre-election debates from 2019 (TVP 2019-10-01 and TVN 2019-10-08)

CLARIN-PL: otwarte narzędzia i aplikacje

- narzędzia **przetwarzania podstawowego** (pełna digitalizacja)
- narzędzia **ekstrakcji informacji z tekstu** (słowa kluczowe, byty nazwane, relacje przestrzenne...)
- narzędzia do **tworzenia i zarządzania korpusami** tekstowymi
- aplikacje badawcze generujące **statystyki tekstów, klasyfikatory tekstów** (treściowe i stylowe)
- narzędzia do **anonimizacji, normalizacji i standaryzacji** tekstów

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.



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)  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



• WageIndicator.org

European *Values* Study



Trudne sprawy...

- dostęp do zdeponowanych publikacji i zasobów naukowych?
- czy można publikować swoje publikacje i zasoby?
- a kto jest za to odpowiedzialny?
- czy muszę się logować?
- czy weryfikacja statusu akademickiego to jeszcze OA?
- czy dane “na prośbę” to jeszcze OA?
- czy udostępnianie moce obliczeniowej?
- czy limitujecie moce obliczeniowej?
- czy promujecie i szkolicie?
- jak finansujecie utrzymanie infrastruktury naukowej?
- otwarci od zawsze?
- czy to działa?

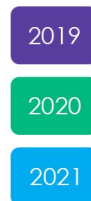
CLARIN-PL Users (VI 2018 – VI 2021)

- Zidentyfikowani użytkownicy danych, narzędzi i aplikacji CLARIN-PL:
 - **199 zespołów i badaczy indywidualnych, w sumie**
 - rodzaj aktywności:
 - od zwykłego używania zasobów i narzędzi językowych, po zaangażowanie zespołu CLARIN-PL (**free!**) w prace nad rozszerzeniem danych, przystosowaniem narzędzi i wykonaniem nowych aplikacji badawczych dla użytkowników
- Spontaniczni użytkownicy: 442
 - wiemy o nich z cytowań i wzmianek w sieci, czasem dzięki kontaktowi osobistemu
- Użytkownicy dydaktyczni: 94
- Statystyki użycia:
 - 6 871 828 zadań przetwarzania tylko w ws.clarin-pl.eu
 - 25 195 294 przetworzonych dokumentów,
 - całkowita objętość : **405 736 [QV = QuoVadis]**
 - 1 [QV] = ilość tekstu w „Quo vadis” Henryka Sienkiewicza

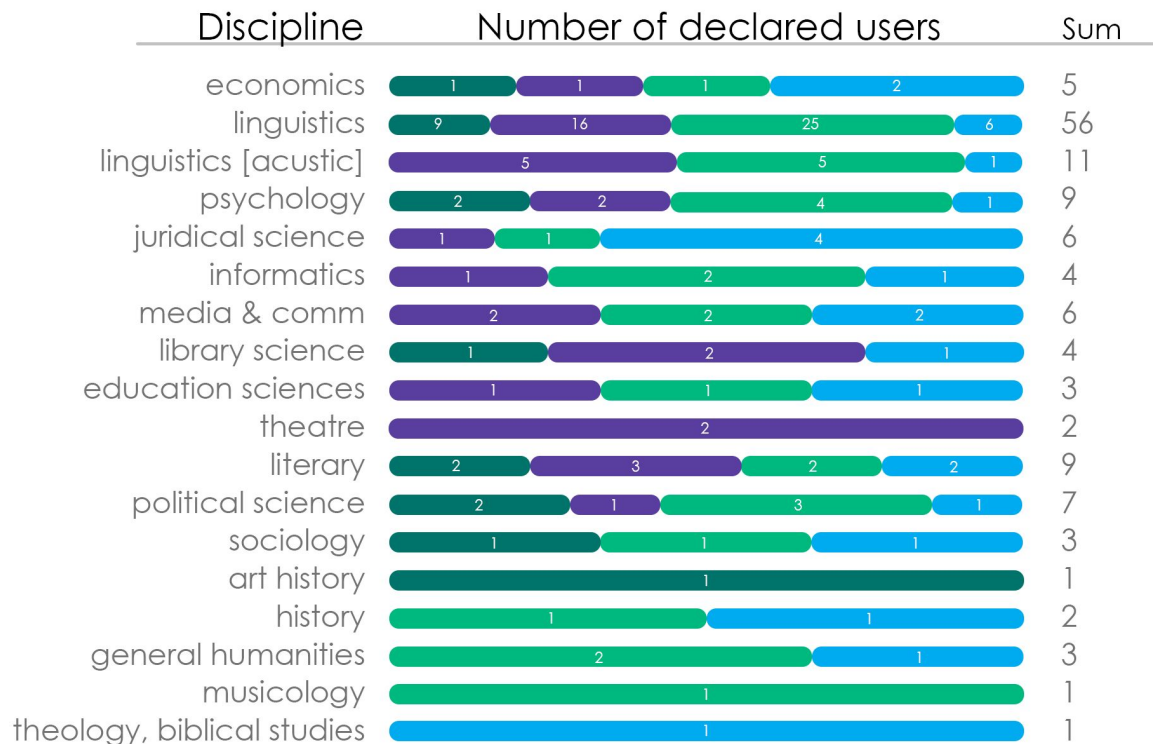
CLARIN-PL Users – citations



Years



CLARIN-PL Users



Years

before
2019

2019

2020

2021

Dziękuję za uwagę!

Więcej informacji nt. CLARIN:

CLARIN-PL
Common Language Resources and Technology Infrastructure



www.clarin.eu
clarin-pl.eu

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

zapraszamy do kontaktu:
clarin@clarin.eu
clarin-pl@pwr.edu.pl

jan.wieczorek@pwr.edu.pl



<http://clarin.eu/>