



Artificial Intelligence in **LUDZIE NAUKI**

Przemysław Korytkowski

Marek Michajłowicz



euroCRIS Strategic Membership Meeting 2024

Paris, 28.11.2024

This work is licensed under CC BY-NC-ND

4.0

Project goals



Promotion and opening up of data on Polish science

- Researchers and their achievements accessible to all
- Contact details for experts



Reliable source of data

- Scientific experts
- Trustworthy data
- Cooperation between science and business



E-services for researchers

- Grants application
- Degree obtaining procedures

Origins of Ludzie Nauki



NAUKA POLSKA

The database of:

- degrees awarded
- disciplines/disciplines
- employment history at Polish institutions
- doctoral thesis supervisors
- doctoral thesis reviews
- history of academic functions



PBN
Polska Bibliografia Naukowa

Collect and present
metadata of
scientific
publications

polon^{2.0}

It collects data as in Nauka Polska and additionally:

- employment history of the scientist in Polish institutions
- patents (national, EPO, PCT)
- projects
- promotions of doctoral theses
- reviews of doctoral theses
- Information on research infrastructure
- ORCID of the researcher



LUDZIE NAUKI
PEOPLE OF SCIENCE

Portal with verified data on scientists and their achievements for:

- researchers
- governmental and local administration
- society
- journalists
- business

which enables the implementation of the State Informatisation Strategy

1999

2013

2014

2023



www

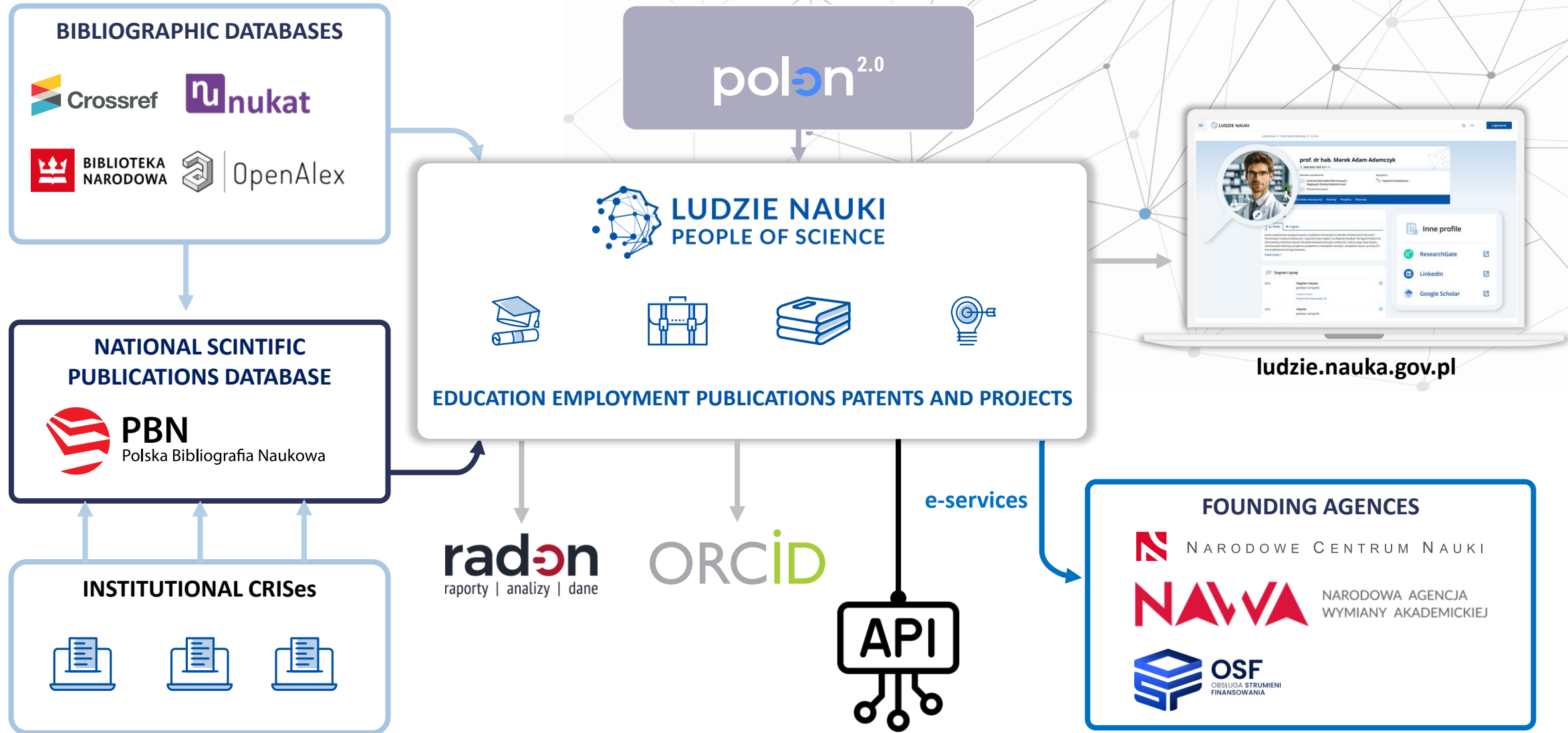


reporting systems



FAIR data system, data can be found, understood and reused

Data workflow



Researchers' profiles



200 000+
profiles



1 700 000+
publications



130 000
artistic achievements



12 000+
patents



36 000+
projects



19 000+
PhD and DSc



prof. dr hab. Marianna Anna Adamczyk

0000-0003-1993-1211

Aktualne zatrudnienie



Centrum Materiałów Polimerowych i
Węglowych Polskiej Akademii Nauk



Politechnika Łódzka

Dyscyplina



inżynieria biomedyczna

O mnie Publikacje Dorobek artystyczny Patenty i prawa ochronne Projekty Prace badawcze



O mnie



Polski



English

[Edytuj informacje](#)

Jestem projektantką oprogramowania i analitykiem biznesowym w Ośrodku Przetwarzania Informacji – Państwowym Instytucie Badawczym. Uzyskałam tytuł magistra na Wydziale Geodezji i Kartografii Politechniki Warszawskiej. Posiadam również kilkuletnie doświadczenie jako analityczka GIS i twórcza usług. Moje obszary zainteresowań obejmują zarządzanie projektami w metodach zwinnych, zarządzanie danymi, procesy ETL oraz projektowanie oprogramowania.

[Pokaż więcej](#)



Dane kontaktowe

[Edytuj w Ustawieniach konta](#)

E-mail

xyz@zyssssz.com

Numer telefonu

[+48 22 570 14 00](tel:+48225701400)



Inne profile

[Edytuj w Ustawieniach konta](#)

Źródło: POL-on



ResearchGate

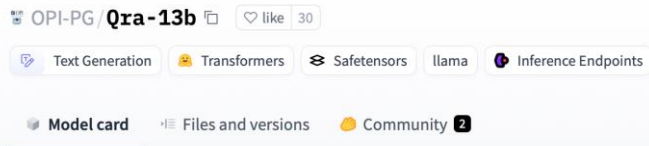


LinkedIn



AI Lab at OPI

- ✓ We trained and fine-tuned many Polish Transformer models:
 - ✓ Collect and pre-process a large corpus of Polish documents (previously 300+ GB raw text, now we have 2+ TB)
 - ✓ Train different transformer-based language models from scratch (e.g. polish RoBERTa, BART, GPT2, Qra)
 - ✓ Train different neural text encoders (e.g. st-paraphrase, mmlw)
 - ✓ Fine-tune the models on a dozen of Polish linguistic tasks (e.g. classification, regression, QA, summarization tasks)
 - ✓ Different evaluations e.g. using Allegro KLEJ benchmark, or ours – PIRB
 - ✓ Release the pre-trained models to the public



Qra-13b



Leaderboard

Rank	Model	Fine-tuning	Average
1	Polish RoBERTa-v2 (large)	Polish RoBERTa scripts	88.9
2	HerBERT (large)	None	88.4
3	XLNet-RoBERTa (large) + NKJP	Polish RoBERTa scripts	87.8
4	Polish RoBERTa (large)	Polish RoBERTa scripts	87.8

Spaces sdadas pibb 15 likes Running

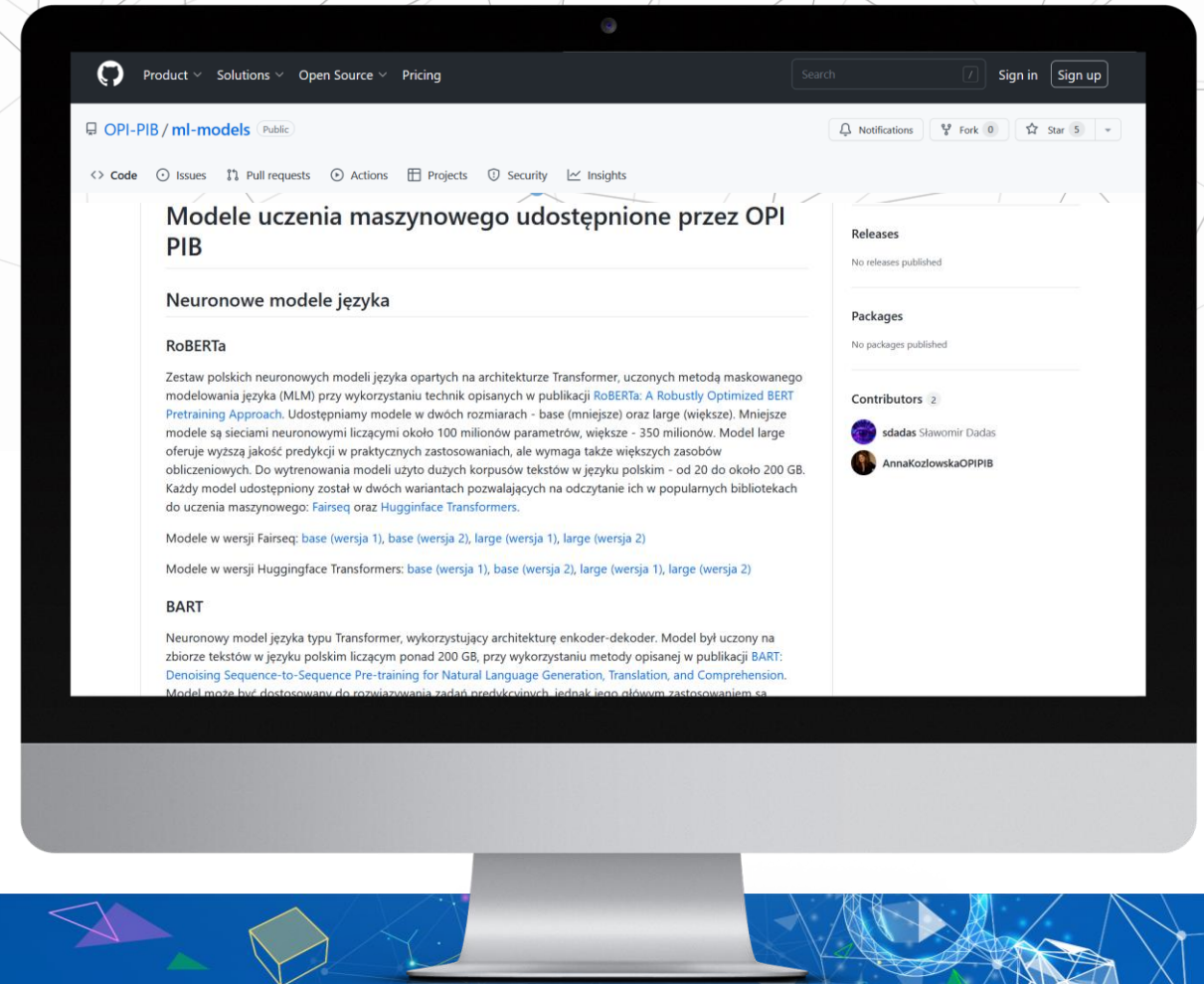
Evaluation metric: NDCG@10 MRR@10 Recall@100 Accuracy@1 Show methods: Retrievers Hybrid retrieval Rerankers

Model	Tasks won	Average (41 tasks)	PolEval-2022 (7 tasks)	Web Datasets (9 tasks)	BEIR-PL (11 tasks)	MaupQA (12 tasks)	Other (2 tasks)
sdadas/mmlw-retrieval-roberta-large	11	58.46	62.72	67.37	53.22	50.20	81.80
sdadas/mmlw-retrieval-e5-large	1	58.30	62.13	66.73	53.53	50.18	81.82
intfloat/multilingual-e5-large	9	57.29	65.86	64.35	48.99	50.53	81.81
sdadas/mmlw-retrieval-roberta-base	0	56.38	59.49	64.52	51.31	49.01	80.98

HOW TO DOWNLOAD OUR Neural Language Models

<https://github.com/OPI-PIB/ml-models>:

- ✓ Pre-trained models, such as RoBERTa, BART, GPT-2, ELMo, Qra
- ✓ Models in two versions:
 - ✓ Pytorch Fairseq
 - ✓ HuggingFace Transformers
- ✓ Evaluation codes are also available
- ✓ Detailed results of the experiments can be found in the corresponding papers



AI in Ludzie Nauki

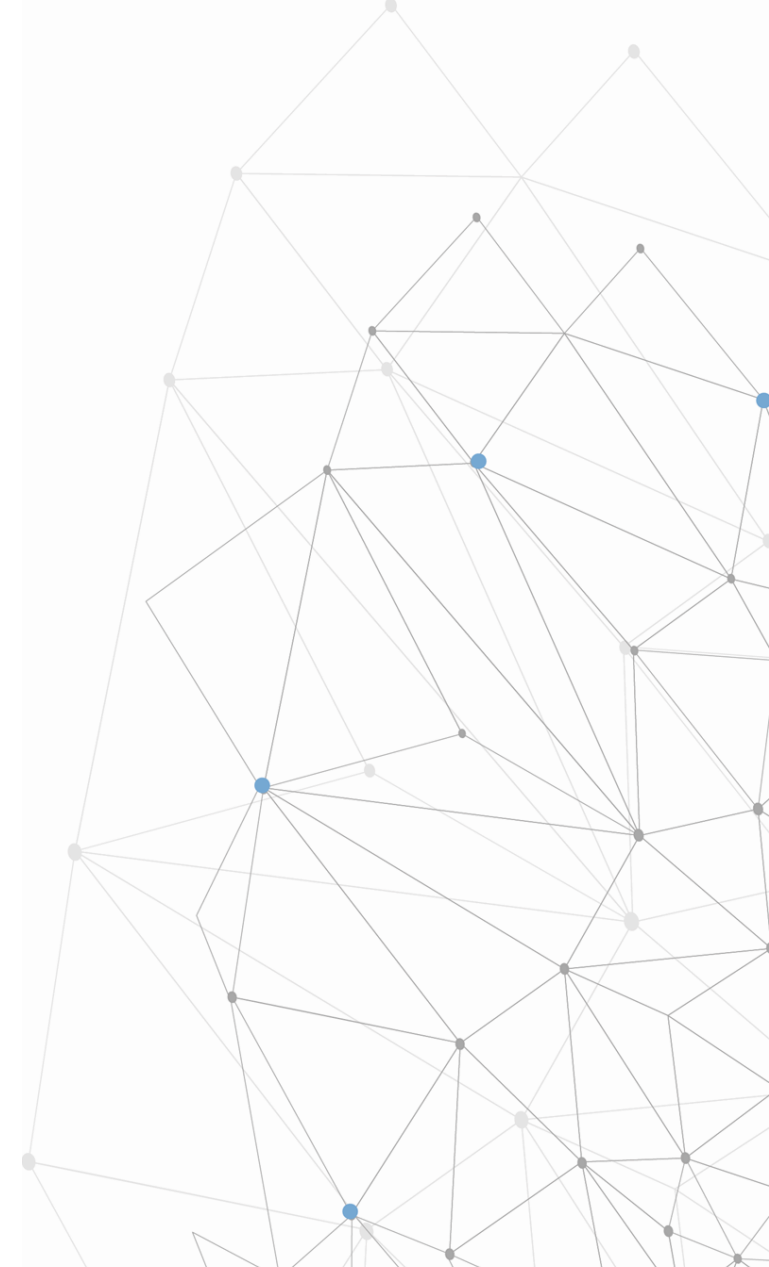
1. Publication deduplication (working)
2. Semantic search engine (under development)
3. Scientific expert search (in the future)



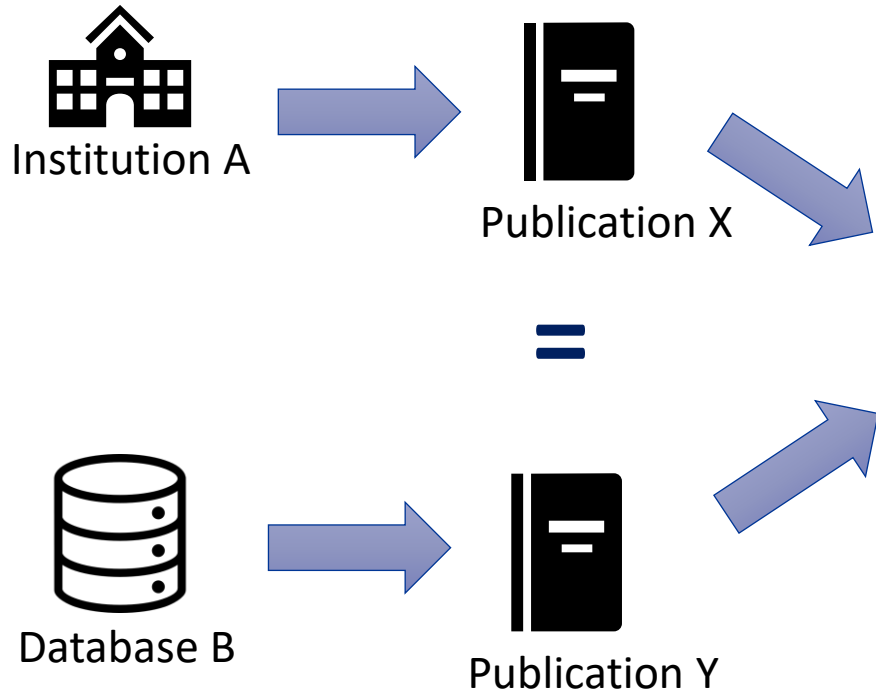
Publication deduplication



The **Golden Publication Record** is the single, richest and most reliable representation of metadata about a scientific publication, at least one of whose authors is an employee of a Polish scientific institution.

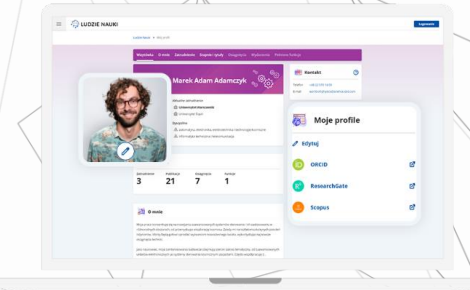
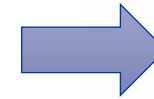


Publication deduplication



Golden Record

1. Publication X from Institution A
2. Publication Y from Database B



Publication X from Institution A
=
Publication Y from Database B

Issues:

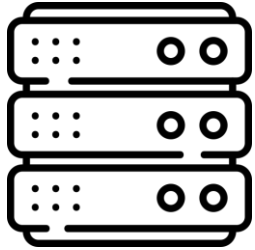
- ✓ Errors in metadata
- ✓ Missing metadata
- ✓ No identifiers for authors outside the institution

Publication deduplication

Standardisation of titles:

- ✓ Replacement of Roman numerals with Arabic numerals,
- ✓ Removal of non-alphanumeric characters (punctuation, inverted commas),
- ✓ Reduction to lower case and replacement of accents with basic characters
- ✓ Removal of HTML and XML tags
- ✓ Title should be longer than 16 characters

Publication deduplication – infrastructure



22 servers (each):

- 16 cores
- 90 GB RAM
- 2 TB disks

352 cores, 1.98TB RAM, 44TB disks



Publication deduplication



PBN
Polska Bibliografia Naukowa

2 103 905



Crossref

343 773

Deduplication

Enrichment

Gold Record

1 774 311

Next sources

 **nukat**



**BIBLIOTEKA
NARODOWA**

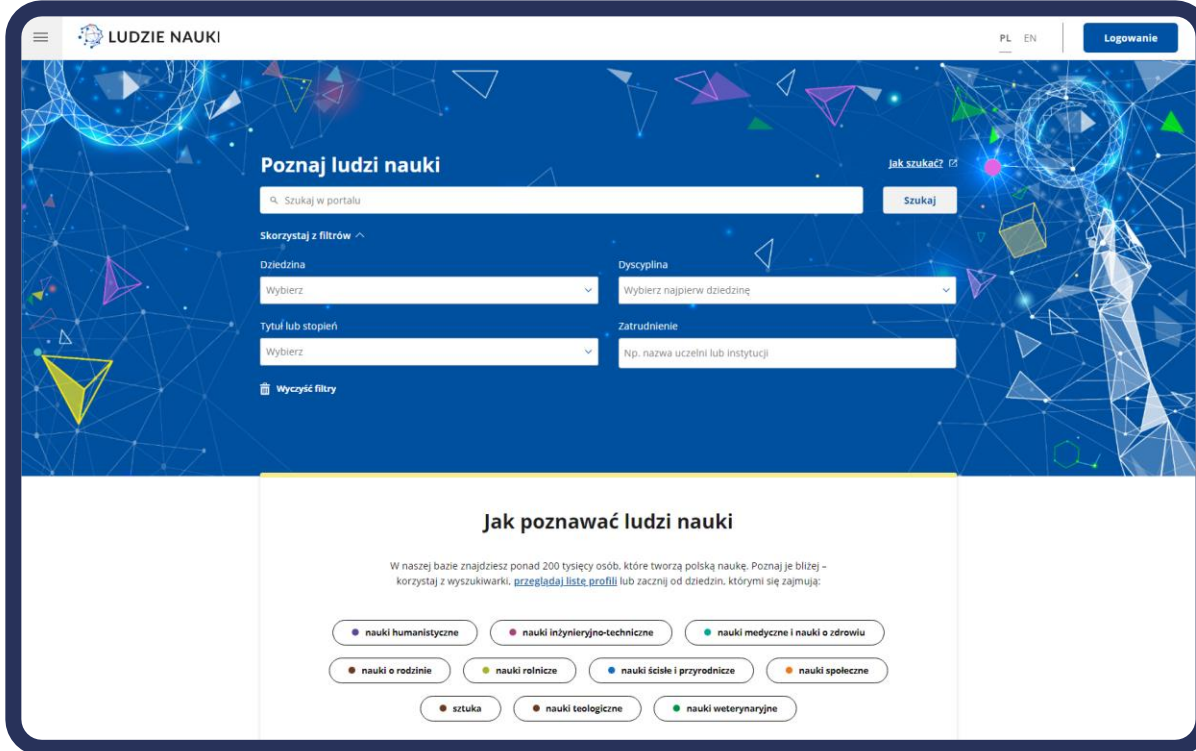


OpenAIRE



OpenAlex

Semantic search engine (under development)



The screenshot shows the 'LUDZIE NAUKI' website interface. At the top, there's a header with the logo, language options (PL, EN), and a 'Logowanie' button. The main section is titled 'Poznaj ludzi nauki' and features a search bar with the placeholder 'Szukaj w portalu' and a 'Szukaj' button. Below the search bar, there are four filter sections: 'Dziedzina' (Field), 'Dyscyplina' (Discipline), 'Tytuł lub stopień' (Title or degree), and 'Zatrudnienie' (Employment). Each filter has a dropdown menu. There's also a 'Wyczyść filtry' button. At the bottom, there's a section titled 'Jak poznawać ludzi nauki' with a description and a grid of category buttons: nauki humanistyczne, nauki inżynierjno-techniczne, nauki medyczne i nauki o zdrowiu, nauki o rodzinie, nauki relnicze, nauki ścisłe i przyrodnicze, nauki społeczne, sztuka, nauki teologiczne, and nauki weterynaryjne.



The goal is to achieve more accurate search results by better understanding user intentions and semantical relations between words in the query.

We are planning to employ Large Language Model based on Transformer architecture with attention mechanism.

Semantic search engine (under development)

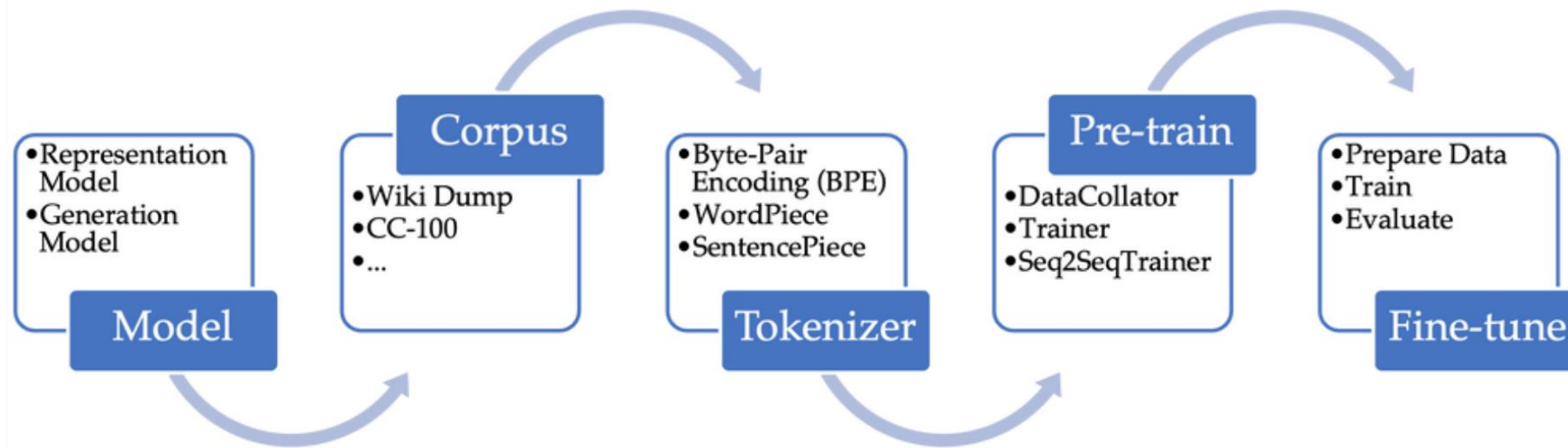


Figure 1. Overview of pre-training and fine-tuning with Hugging Face.

Scientific expert search (in the future)



Efficient selection of experts

- shorter selection times for experts
- increased availability of experts
- increasing the number of expert candidates



Increasing the transparency of the expert selection process

- Reducing subjectivity in the expert selection process



Efficient recruitment of domain experts

- Reducing the time and cost of recruiting and informing experts about the process



OŚRODEK
PRZETWARZANIA
INFORMACJI
PAŃSTWOWY INSTYTUT BADAWCZY

Thank You for Your attention!
Dziękuję za uwagę!