Sztuczna inteligencja w rozwoju bibliograficznych baz danych

dr hab. inż. Przemysław Korytkowski, prof. ZUT, prof. OPI-PIB

Zachodniopomorski Uniwersytet Technologiczny w Szczecinie
Ośrodek Przetwarzania Informacji – Państwowy Instytut Badawczy

Lublin, 27 września 2024

1

Ekstrakcja cytowań

Projekt “System informatyczny do analizy bibliografii polskich tekstów naukowych”

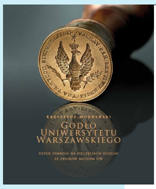
Finansowanie z program Nauka dla społeczeństwa II, MNiSW

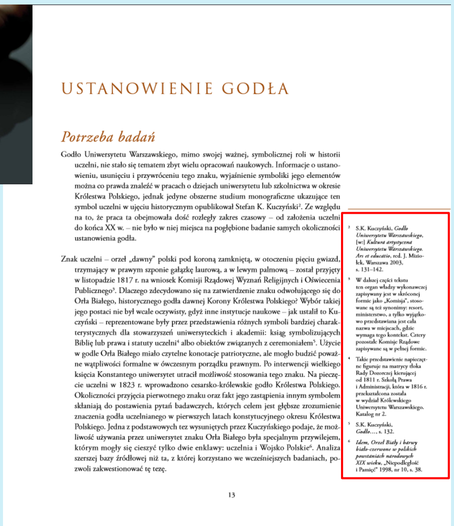
Okres realizacji: luty 2025 – luty 2028

Wykonawcy: ZUT i OPI-PIB

2

Problem badawczy





USTANOWIENIE GODŁA

Potrzeba badań

Godło Uniwersytetu Warszawskiego, mimo swojej ważnej, symbolicznej roli w historii uczelni, nie stało się tematem zbyt wielu opracowań naukowych. Informacje o ustanowieniu, usunięciu i przywróceniu tego znaku, wyjaśnienie symboliki jego elementów można co prawda znaleźć w pracach o dziejach uniwersytetu lub szkolnictwa w okresie Królestwa Polskiego, jednak jedyne obszerne studium monograficzne ukazujące ten symbol uczelni w ujęciu historycznym opublikował Stefan K. Kuczyński[2]. Ze względu na to, że praca ta obejmowała dość rozległy zakres czasowy – od założenia uczelni do końca XX w. – nie było w niej miejsca na pogłębione badanie samych okoliczności ustanowienia godła.

Znak uczelni – orzeł „dawny" polski pod koroną zamkniętą, w otoczeniu pięciu gwiazd, trzymający w prawym szponie gałązkę laurową, a w lewym palmową – został przyjęty w listopadzie 1817 r. na wniosek Komisji Rządowej Wyznań Religijnych i Oświecenia Publicznego[3]. Dlaczego zdecydowano się na zatwierdzenie znaku odwołującego się do Orła Białego, historycznego godła dawnej Korony Królestwa Polskiego? Wybór takiej jego postaci nie był wcale oczywisty, gdyż inne instytucje naukowe – jak ustalił to Kuczyński – reprezentowane były przez przedstawienia różnych symboli bardziej charakterystycznych dla stowarzyszeń uniwersyteckich i akademii: ksiąg symbolizujących Biblię lub prawa i statuty uczelni[4] albo obiektów związanych z ceremoniałem[5]. Użycie w godle Orła Białego miało czytelne konotacje patriotyczne, ale mogło budzić poważne wątpliwości formalne w ówczesnym porządku prawnym. Po interwencji wielkiego księcia Konstantego uniwersytet utracił możliwość stosowania tego znaku. Na pieczęcie uczelni w 1823 r. wprowadzono cesarsko-królewskie godło Królestwa Polskiego. Okoliczności przyjęcia pierwotnego znaku oraz fakt jego zastąpienia innym symbolem skłaniają do postawienia pytań badawczych, których celem jest głębsze zrozumienie znaczenia godła uczelnianego w pierwszych latach konstytucyjnego okresu Królestwa Polskiego. Jedna z podstawowych tez wysuniętych przez Kuczyńskiego podaje, że możliwość używania przez uniwersytet znaku Orła Białego była specjalnym przywilejem, którym mogły się cieszyć tylko dwie enklawy: uczelnia i Wojsko Polskie[6]. Analiza szerszej bazy źródłowej niż ta, z której korzystano we wcześniejszych badaniach, pozwoli zakwestionować tę tezę.

[2] S.K. Kuczyński, *Godło Uniwersytetu Warszawskiego*, [w:] *Kultura artystyczna Uniwersytetu Warszawskiego. Ars et educatio*, red. J. Miziołek, Warszawa 2003, s. 131–142.

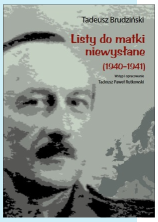
[3] W dalszej części tekstu ten organ władzy wykonawczej zapisywany jest w skróconej formie jako „Komisja", stosowane są też synonimy: resort, ministerstwo, a tylko wyjątkowo przedstawiana jest cała nazwa w miejscach, gdzie wymaga tego kontekst. Cztery pozostałe Komisje Rządowe zapisywane są w pełnej formie.

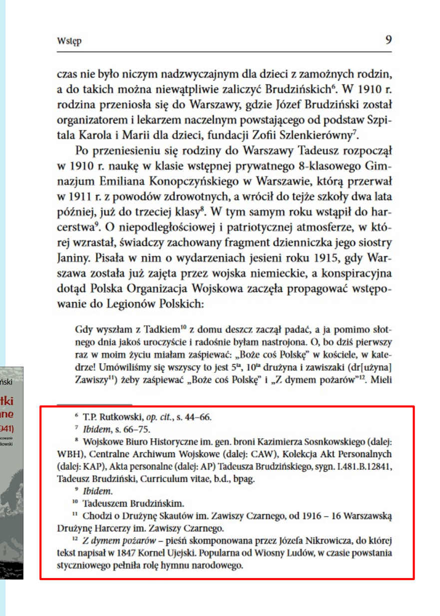
[4] Takie przedstawienie napieczętne figuruje na matrycy tłoka Rady Dozorczej kierującej od 1811 r. Szkołą Prawa i Administracji, która w 1816 r. przekształcona została w wydział Królewskiego Uniwersytetu Warszawskiego. Katalog nr 2.

[5] S.K. Kuczyński, *Godło…*, s. 132.

[6] *Idem, Orzeł Biały i barwy biało-czerwone w polskich powstaniach narodowych XIX wieku*, „Niepodległość i Pamięć" 1998, nr 10, s. 38.

13





Wstęp 9

czas nie było niczym nadzwyczajnym dla dzieci z zamożnych rodzin, a do takich można niewątpliwie zaliczyć Brudzińskich[6]. W 1910 r. rodzina przeniosła się do Warszawy, gdzie Józef Brudziński został organizatorem i lekarzem naczelnym powstającego od podstaw Szpitala Karola i Marii dla dzieci, fundacji Zofii Szlenkierówny[7].

Po przeniesieniu się rodziny do Warszawy Tadeusz rozpoczął w 1910 r. naukę w klasie wstępnej prywatnego 8-klasowego Gimnazjum Emiliana Konopczyńskiego w Warszawie, którą przerwał w 1911 r. z powodów zdrowotnych, a wrócił do tejże szkoły dwa lata później, już do trzeciej klasy[8]. W tym samym roku wstąpił do harcerstwa[9]. O niepodległościowej i patriotycznej atmosferze, w której wzrastał, świadczy zachowany fragment dzienniczka jego siostry Janiny. Pisała w nim o wydarzeniach jesieni roku 1915, gdy Warszawa została już zajęta przez wojska niemieckie, a konspiracyjna dotąd Polska Organizacja Wojskowa zaczęła propagować wstępowanie do Legionów Polskich:

> Gdy wyszłam z Tadkiem[10] z domu deszcz zaczął padać, a ja pomimo słotnego dnia jakoś uroczyście i radośnie byłam nastrojona. O, bo dziś pierwszy raz w moim życiu miałam zaśpiewać: „Boże coś Polskę" w kościele, w katedrze! Umówiliśmy się wszyscy to jest 5ta, 10ta drużyna i zawiszaki (dr[użyna] Zawiszy[11]) żeby zaśpiewać „Boże coś Polskę" i „Z dymem pożarów"[12]. Mieli

[6] T.P. Rutkowski, *op. cit.*, s. 44–66.

[7] *Ibidem*, s. 66–75.

[8] Wojskowe Biuro Historyczne im. gen. broni Kazimierza Sosnkowskiego (dalej: WBH), Centralne Archiwum Wojskowe (dalej: CAW), Kolekcja Akt Personalnych (dalej: KAP), Akta personalne (dalej: AP) Tadeusza Brudzińskiego, sygn. I.481.B.12841, Tadeusz Brudziński, Curriculum vitae, b.d., bpag.

[9] *Ibidem*.

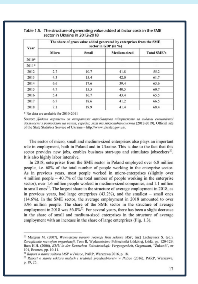
[10] Tadeuszem Brudzińskim.

[11] Chodzi o Drużynę Skautów im. Zawiszy Czarnego, od 1916 – 16 Warszawską Drużynę Harcerzy im. Zawiszy Czarnego.

[12] *Z dymem pożarów* – pieśń skomponowana przez Józefa Nikrowicza, do której tekst napisał w 1847 Kornel Ujejski. Popularna od Wiosny Ludów, w czasie powstania styczniowego pełniła rolę hymnu narodowego.

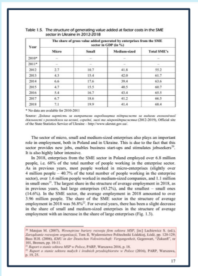
3

Lokalizacja przypisu



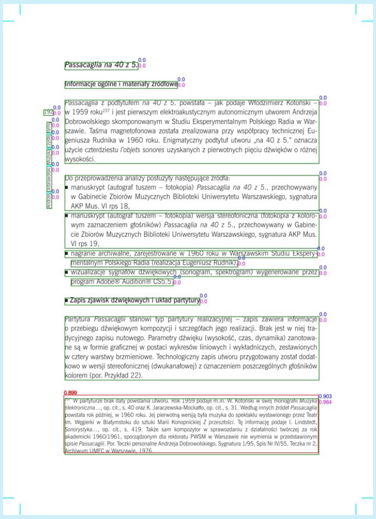
Obraz

Detekcja obiektów



Współrzędne przypisu na obrazie

Tekst



4

Przetworzenie przypisu

87 Por. C. Sterling, Sieć terroru, Warszawa 1990.

Tekst

NER

87 Por. C. Sterling , Sieć terroru , Warszawa 1990 .

Wykryte podmioty

98 M. Grzechnik , ' Ad Maiorem Poloniae Gloriam!' Polish Inter-colonial Encounters in Africa in the Interwar Period , „ The Journal of Imperial and Commonwealth History " 2020 , nr 48 (5), s. 826–845 ; B. Balogun , Polish Lebensraum: the colonial ambition to expand on racial terms , „ Ethnic and Racial Studies " 2018 , nr 41 (14), s. 2561–2579 .

5

Identyfikacja i wzbogacanie

64 K.W. Mucha, Obóz narodowy..., s. 297–298.

System ekstrakcji cytowań

Wzbogacanie danych

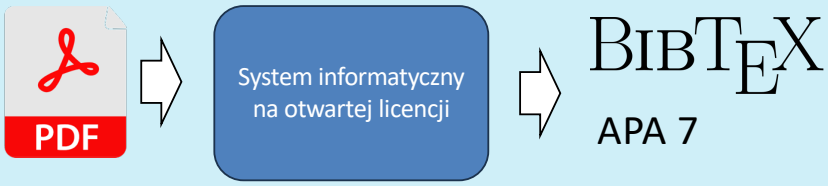
BIBLIOTEKA NARODOWA

PBN Polska Bibliografia Naukowa

Crossref

6

Wynik projektu

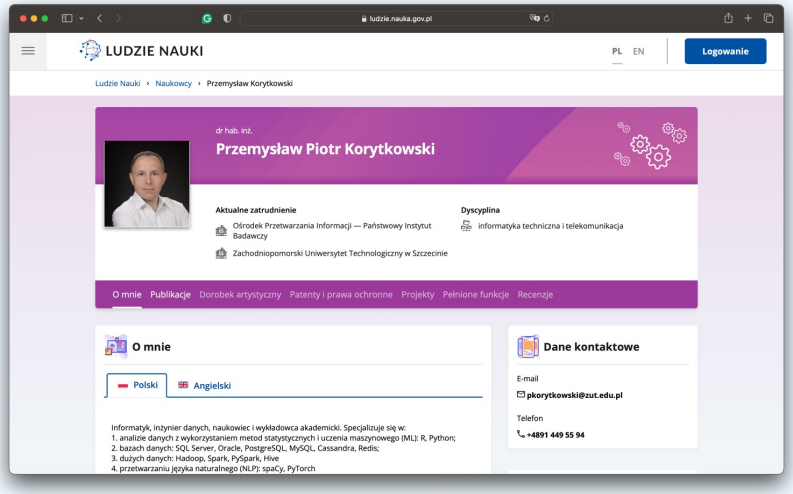


7

Ludzie Nauki

ludzie.nauka.gov.pl

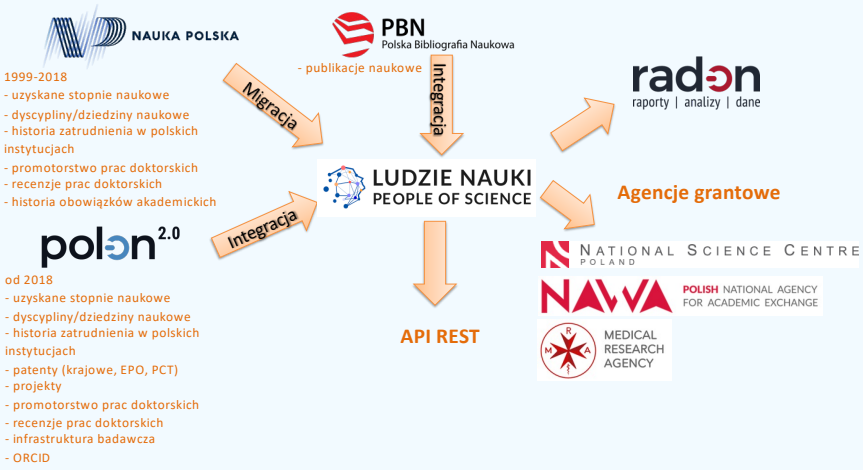
200 000+ profili naukowców
2 000 000+ metadanych publikacji
130 000 osiągnięć artystycznych
12 000+ patentów
36 000+ projektów



Państwowe dane = wiarygodne dane

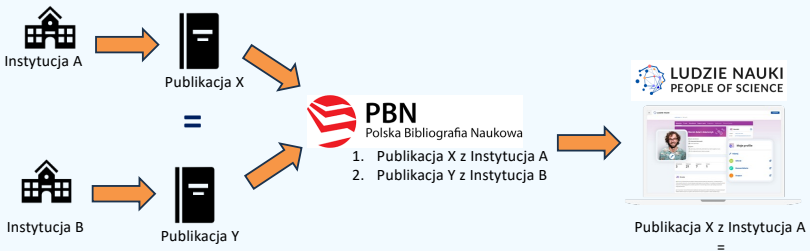
8

Docelowy ecosystem Ludzi Nauki



9

Jakość danych bibliograficznych



Problemy:

1. Błędy w metadanych
2. Braki w metadanych
3. Brak identyfikatorów autorów spoza instytucji

10

Złoty Rekord Publikacji

Złoty rekord to termin używany w dziedzinie zarządzania danymi i odnosi się do pojedynczej i wiarygodnej reprezentacji danych, która pochodzi z różnych źródeł i jest przechowywana w centralnej bazie danych.

Złoty Rekord Publikacji to pojedyncza, najbogatsza i wiarygodna reprezentacja metadanych o publikacji naukowej, której przynajmniej jednym z autorów jest pracownik polskiej instytucji naukowej.

11

Złoty Rekord Publikacji

1. **Gromadzenie danych** z różnych źródeł w celu stworzenia kompleksowego obrazu podmiotu danych.
2. **Czyszczenie i standaryzacja** danych, poprawianie błędów i standaryzacja danych w celu zapewnienia ich spójności i dokładności.
3. **Usuwanie duplikatów**, identyfikacja i łączenie duplikatów i sprzecznych informacji w celu stworzenia pojedynczej, wiarygodnej reprezentacji metadanych.
4. **Wzbogacanie danych**, dodawanie brakujących lub uzupełniających informacji do złotego rekordu, aby uczynić go bardziej kompletnym i użytecznym.
5. **Walidacja danych**, sprawdzanie dokładności złotego rekordu i upewnianie się, że jest on zgodny z regułami biznesowymi i standardami jakości danych.

12

Deduplikacja metadanych publikacji

13

Złoty Rekord Publikacji

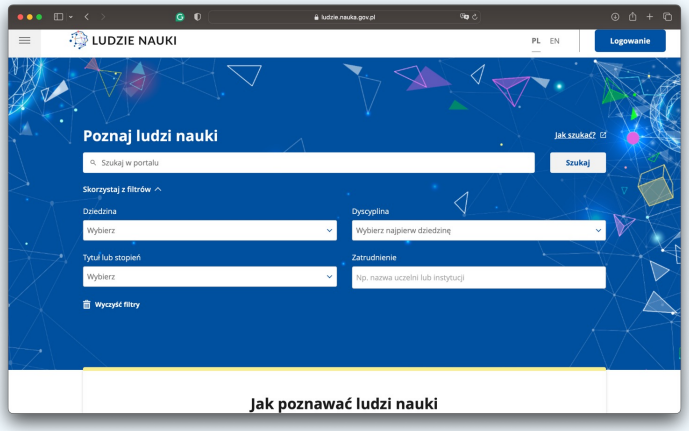
1. Wstępne wyszukiwanie par tytułów z wykorzystaniem algorytmów TF oraz Hierarchical Navigable Small World
2. Zgodność musi być na poziomie 0,9 odległości cosinusowej
3. Wartości następujących atrybutów muszą się zgadzać (te same wartości lub NULL): DOI, ISBN, ISSN, ISMN, rok, MNISW_ID
4. Jeśli w tytule cyfry się nie zgadzają, to para publikacji nie jest łączona (np. tom 1 i tom II)
5. Tytuł musi być dłuższy niż 16 znaków
6. Nazwa czasopisma i nazwa wydawnictwa musi być zgodne na poziomie 0,8 odległości cosinusowej
7. Tytuły książki w przypadku rozdziałów muszą być zgodne na poziomie 0,9 odległości cosinusowej

14

Plany – wyszukiwarka semantyczna

Wyszukiwanie semantyczne to podejście, które stara się uzyskać jak najdokładniejsze wyniki poprzez zrozumienie intencji użytkownika oraz powiązań między słowami i kontekstem samego zapytania.

Planujemy wykorzystać duże modele językowe (LLM).



15

Konkluzje

Połączone Otwarte Dane

Dla publikacji naukowych:

1. DOI – Digital Object Identifier
2. ORCID – Open Researcher and Contributor ID
3. ROR – Research Organization Registry

By Michael Hausenblas, James G. Kim, five-star Linked Open Data rating system developed by Tim Berners-Lee. - http://5stardata.info/en/, CC0, https://commons.wikimedia.org/w/index.php?curid=64408651

16

Dziękuję za uwagę!

dr hab. inż. Przemysław Korytkowski, prof. ZUT, prof. OPI-PIB

Zachodniopomorski Uniwersytet Technologiczny w Szczecinie
Ośrodek Przetwarzania Informacji – Państwowy Instytut Badawczy

Lublin, 27 września 2024

17