



ADAM MICKIEWICZ UNIVERSITY IN POZNAŃ
Scholarly Communication Research Group

Budowa i przyszłość Korpusu Polskich Czasopism Naukownawczych



@ekulczycki

Emanuel Kulczycki

Kulczycki, E., Zambrano Mena, Y. A., & Krawczyk, F.
(2023). **Budowa i charakterystyka Korpusu
Polskich Czasopism Naukowych.**
Zagadnienia Informacji Naukowej, 61(2), 9–31.
<https://doi.org/10.36702/zin2023.02.01>

**Budowa i charakterystyka
Korpusu Polskich
Czasopism Naukowych**

Emanuel Kulczycki*,
ORCID: 0000-0001-6530-3609

Yeimer Alexander Zambrano Mena,

Franciszek Krawczyk
ORCID: 0000-0003-1097-9032

*Pracownia Komunikacji Naukowej
Uniwersytet im. Adama Mickiewicza w Poznaniu*

Abstrakt

Cel/teza: Artykuł przedstawia Korpus Polskich Czasopism Naukowych (KPCN), to jest specjalistyczny korpus stworzony w celu wsparcia badań w dziedzinie naukoznawstwa oraz jego rozwoju w Polsce.

Koncepcja/Metody badań: Budowa korpusu oparta była na digitalizacji wcześniej nie-digitalizowanych artykułów oraz pobieraniu tekstów ze stron internetowych czasopism naukowych i bibliotek cyfrowych, które zostały poddane metodom przetwarzania języka naturalnego.

Wyniki i wnioski: Możliwości KPCN zademonstrowano poprzez analizę modelowania tematycznego czasopisma „Nauka Polska”. Obecna wersja KPCN obejmuje 12 polskich czasopism naukowych z lat 1918–2020, zawierających łącznie 51 822 dokumenty.

Ograniczenia badań: Badanie uznaje pewne ograniczenia korpusu, zwłaszcza w kontekście przetwarzania języka naturalnego i optycznego rozpoznawania tekstu. Pomimo zauważonych ograniczeń, artykuł bada również możliwości przyszłego rozwoju korpusu.

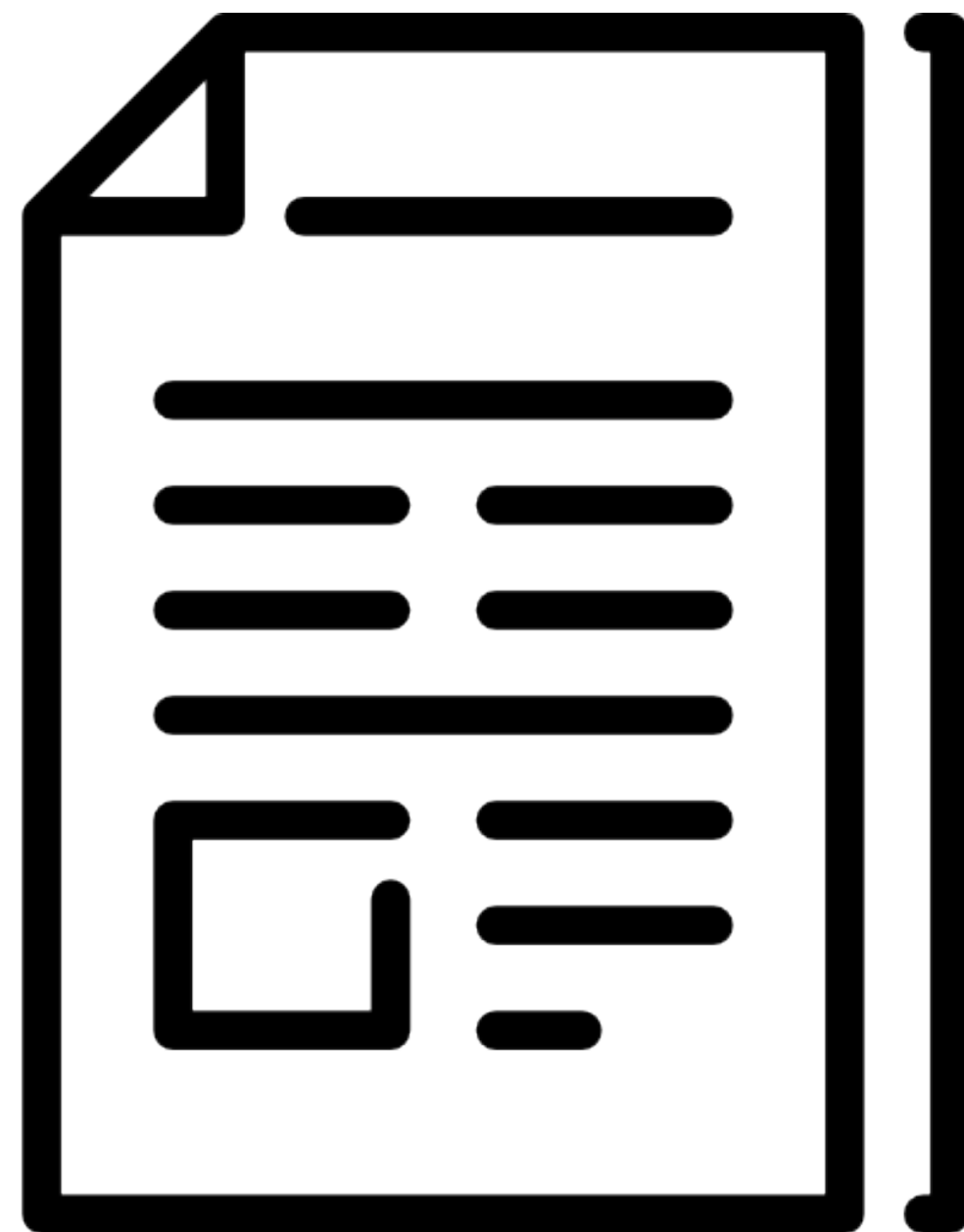
Zastosowania praktyczne: W przyszłości korpus może ułatwić rekonstrukcję dyskursów związanych z nauką i szkolnictwem wyższym w Polsce, przyczyniając się do zwiększenia rozpoznawalności polskiego naukoznawstwa na arenie międzynarodowej.

Oryginalność/wartość: Budowa tego korpusu stanowi oryginalne przedsięwzięcie, obejmujące digitalizację i przetwarzanie artykułów naukowych z dziedziny naukoznawstwa. Ten wysiłek zaowocował stworzeniem unikatowego narzędzia do analizy dyskursów.

Słowa kluczowe

Bibliometria. Korpus tematyczny. Modelowanie tematyczne. Naukoznawstwo. Polska nauka.

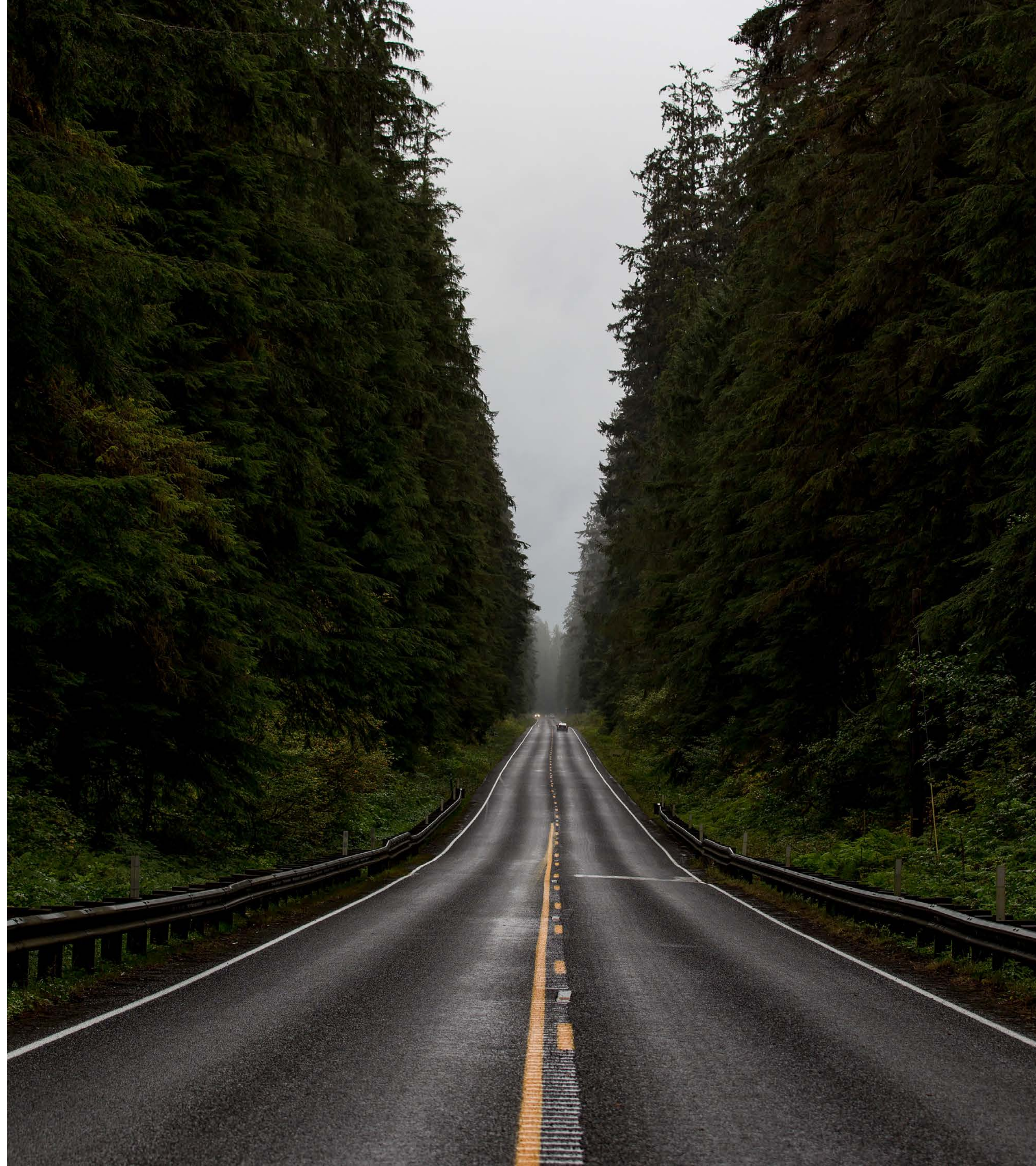
Tekst wpłynął do Redakcji: 2 listopada 2023 r.



Korpus Polskich Czasopism Naukowych jest korpusem tematycznym (ang. *topic-specific corpus*), czyli cyfrowym zbiorem dokumentów pochodzących z polskich czasopism naukowych, który ma służyć badaniom naukowym oraz badaniom nad szkolnictwem wyższym i nauką.

Po co nam korpus?

- ☑ Polska ma unikatowy dorobek naukowawczy, oparty głównie na czasopismach. Ten dorobek jest szerzej nieznany na świecie, a w Polsce przebadany wyrywkowo i często jednostronnie.
- ☑ Większość dorobku czasopiśmienniczego nie jest zdigitalizowana.
- ☑ Dorobek ten jest ogromny, nie do „strawienia przez pojedynczego badacza”.
- ☑ Dysponujemy narzędziami, które pozwalają wygodniej korzystać z dużych zasobów tekstowych.



Polskie naukoznawstwo

- ☑️ Polskie naukoznawstwo (nauka o nauce) jest sprawadzane do dorobku i działalności Kasy Mianowskiego (Florian Znaniecki oraz Maria i Stanisław Ossowsky, Tadeusz Kotarbiński, Kazimierz Ajdukiewicz, Izydora Dąmbska).
- ☑️ Polska „nauka o nauce” jest niekiedy mitologizowana w Polsce i pomija się powojenną dyskusję naukoznawczą/naukometryczną, gdy nie była kontynuacją myśli powstałych wokół Kasy Mianowskiego i czasopisma „Nauka Polska. Jej Potrzeby, Organizacja i Rozwój”.

Dominujące, mitologizujące, podejście

Bardzo cenne prace, lecz lecz niekiedy stronnice w antykomunistycznym i antysocjalistycznym odczytaniu polityki naukowej.

- ☑ *Encyklopedia polskiej nauki akademickiej oraz Polityka naukowa w Polsce w latach 1944-1953*
prof. Piotra Hübnera
- ☑ Publikacje Jana Kozłowskiego.

**Dzisiejsze wyzwania, czyli
stare przeboje w nowym
aranżacjach**

Radca ministerstwa na posiedzeniu Koła Naukoznawczego w 1929 r. wygłosił referat „O stosunku państwa do nauki”

wyższej twórczości (np. przez ustanowienie nagrody państwowej za prace naukową wyjątkowego znaczenia). Podstawą działalności Wydziału Nauki w zakresie dobrze skierowanego rozdziału środków materialnych winny być krytyczne oceny dorobku naukowego poszczególnych instytucji naukowych; państwo jako dawca dóbr materialnych powinno domagać się opracowania ich przez specjalistów. W Państwie Polskiem ingerencja czynników państwowych w sprawach nauki jest większa niż w państwach starych i bogatych, gdzie instytucje spo-

Sprawozdanie czwarte z działalności Koła Naukoznawczego (1932)

26.I.1932 referat dra T. Makowieckiego p. t. «W sprawie ustalenia programu badań w dziedzinie kultury» (12,5).

Po pierwszym odczycie, a z kolei w Kole Naukoznawczem, wywiązała się dyskusja, która dotykała wielu dziedzin pokrewnych. M. in. zwrócono uwagę na ważność obliczeń statystycznych z zakresu produkcji naukowej, gdyż uwydatniają one istotny wkład nauki narodowej w każdej dziedzinie w sposób bardziej wierny, niż dane dotyczące liczby specjalistów, ich udziału w akademjach narodowych, liczby towarzystw naukowych i t. p. Tego rodzaju obliczenia co do nauki polskiej, choć bardzo mozolne, byłyby najlepszym przyczynkiem do charakterystyki jej stanowiska w nauce wszechświatowej.—Da-

Uwagi o organizacji twórczości i pracy naukowo-badawczej w Polsce

Wojciech Świętosławski, 1929

Ważnym bardzo czynnikiem byłoby także utworzenie *ciała kontrolującego*, stwierdzającego co pewien okres czasu *stopień wyzyskania*, a więc i *owocność wysiłku*, *dokonanego w ośrodkach pracy badawczej, naukowej i wynalazczej*. Sprawa ta wymagałaby, oczywiście, bardzo subtelного traktowania, jednakże rozwiązanie pomyślne tego zagadnienia posiada dla organizacji twórczości w państwie znaczenie pierwszorzędne.

Budowanie korpusu

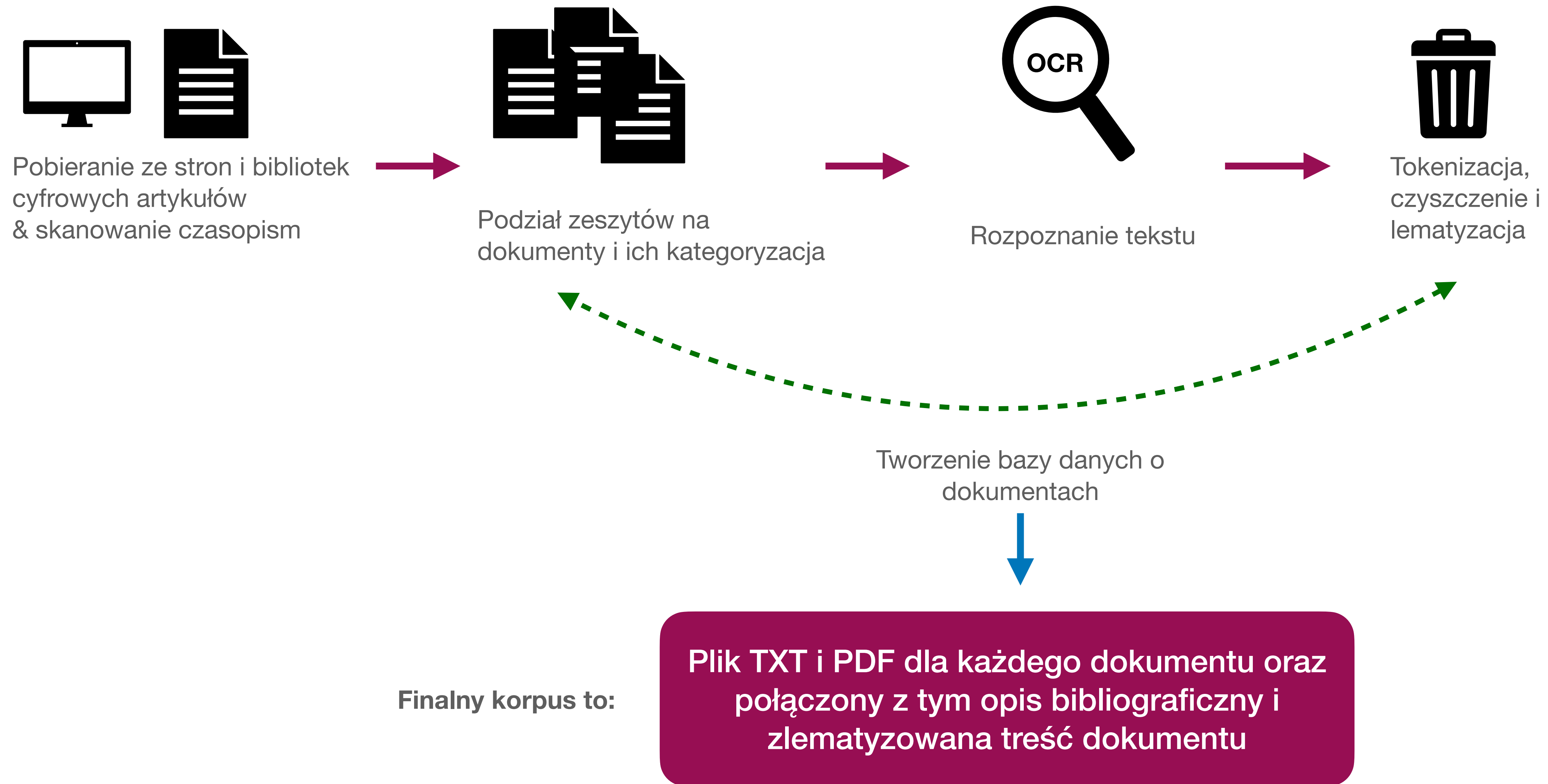
Emanuel Kulczycki
Yeimer Alexander Zambrano Mena
Franciszek Krawczyk

oraz wspaniała armia bibliotekarzy

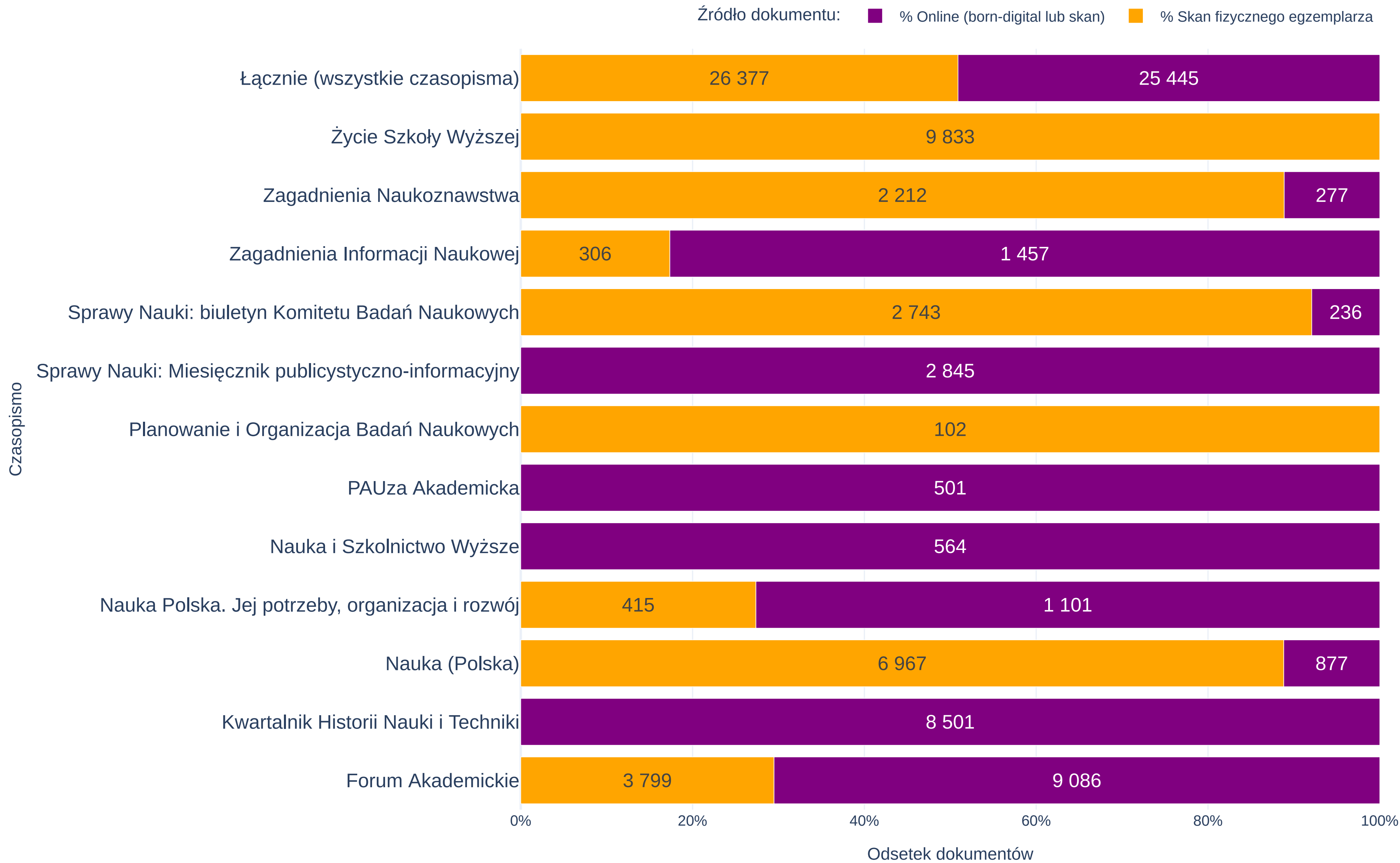
Korpus Polskich Czasopism Naukowych

Lp.	Czasopismo (ostatni tytuł)	Lata ukazywania się (do 2020)	Liczba dokumentów
1	„Forum Akademickie”	1991–2020	12 885
2	„Kwartalnik Historii Nauki i Techniki”	1956–2020	8 496
3	„Nauka”	1954–2020	7 844
4	„Nauka i Szkolnictwo Wyższe”	1993–2019	1 516
5	„Nauka Polska. Jej Potrzeby Organizacja i Rozwój”	1918–1920, 1923, 1925, 1927–1939, 1947, 1992–2020	564
6	„PAUza Akademicka”	2008–2020	501 (liczba zeszytów)
7	„Planowanie i Organizacja Badań Naukowych”	1980, 1982–1987, 1989	102
8	„Sprawy Nauki: biuletyn Komitetu Badań Naukowych”	1991–2009	2 845
9	„Sprawy Nauki: Miesięcznik Publicystyczny-Informacyjny”	2006–2020	2 979
10	„Zagadnienia Informacji Naukowej”	1962–2020	1 763
11	„Zagadnienia Naukoznawstwa”	1965–2019	2 484
12	„Życie Szkoły Wyższej”	1946–1952; 1953–1991	9 833
	Łącznie		51 822

Workflow budowy korpusu



Źródło dokumentów w Korpusie Polskich Czasopism Naukowych



Kategoryzacja dokumentów

Jako dokumenty **niew**chodzące do analizy uznaliśmy:

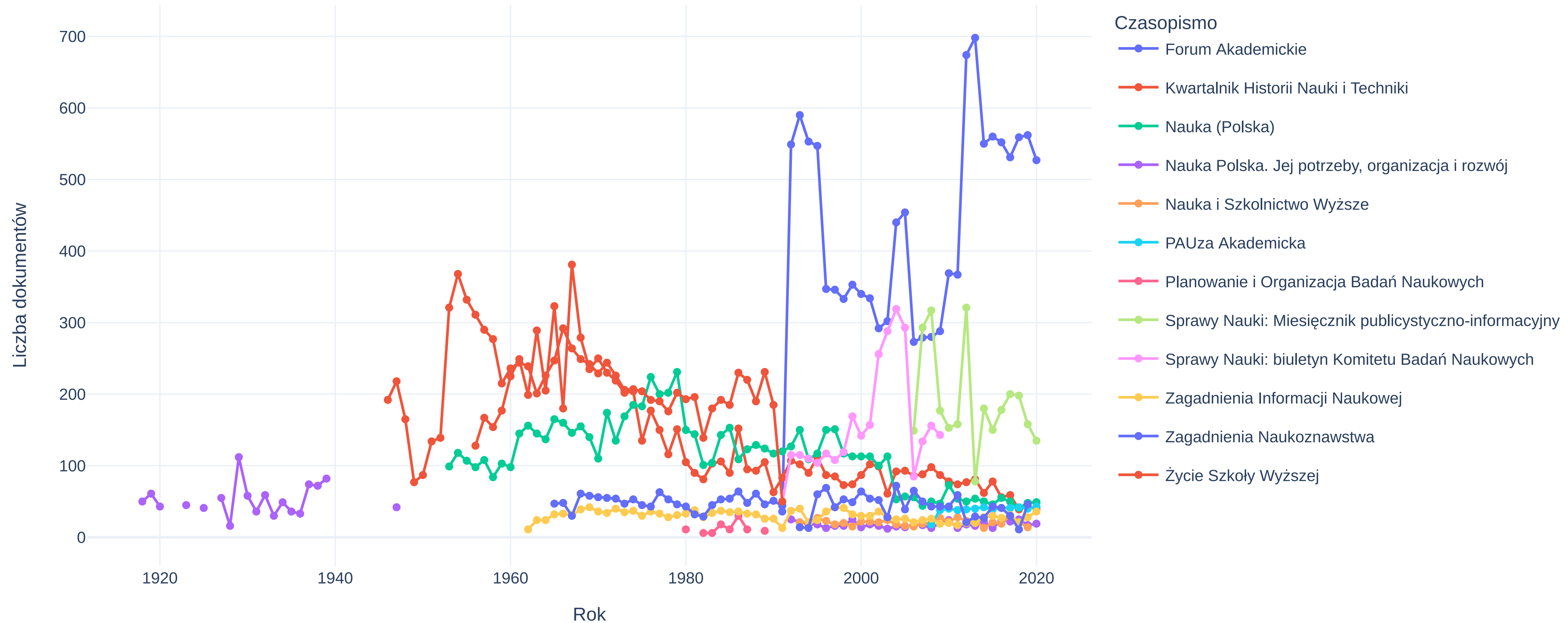
- Strony redakcyjne
- Spisy treści
- Biografie
- Dokumenty składają się jedynie z tabel lub zestawień numerycznych.
- Zestawienia bibliograficzne
- Ogłoszenia
- Streszczenia artykułów
- Artykuły niepolskojęzyczne

Wszystkie dokumenty z korpusu były poddawane tym samym działaniom (tj. rozpoznanie tekstów, przetwarzanie tekstu).

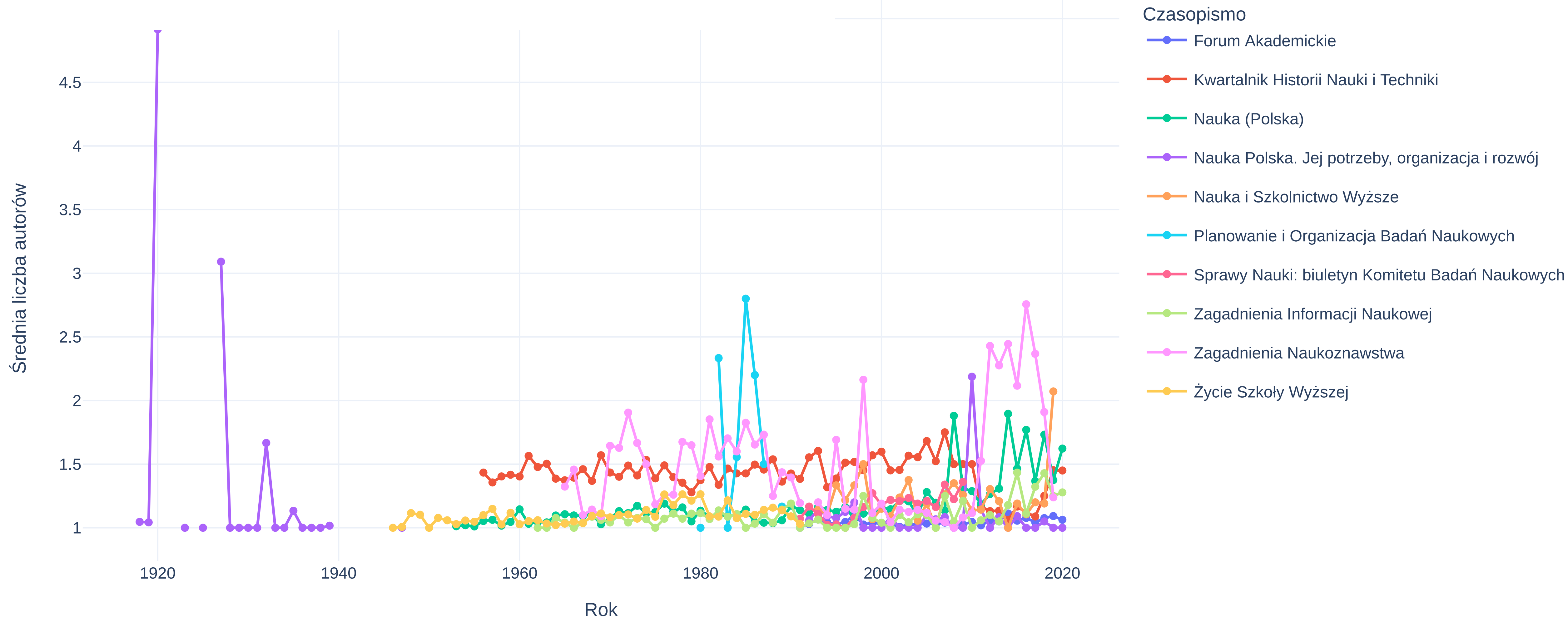


Podstawowa charakterystyka

Liczba dokumentów

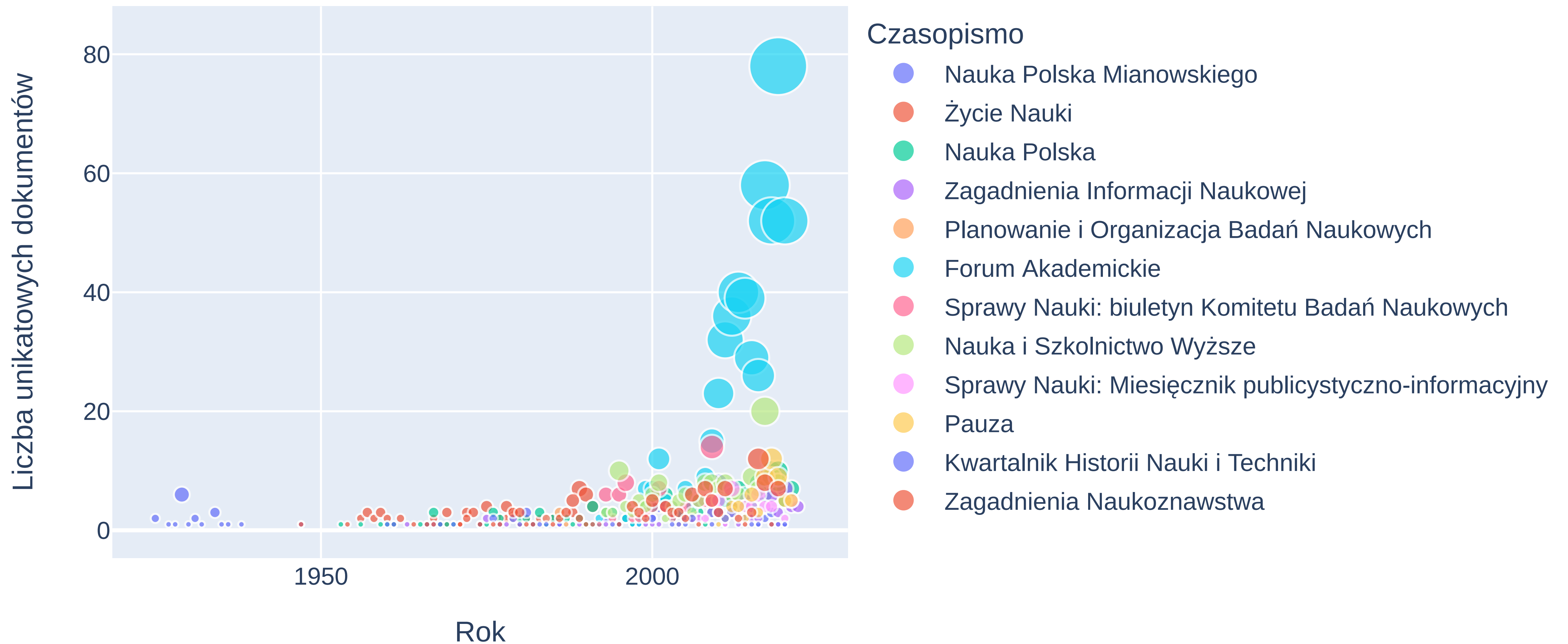


Średnia liczba autorów dokumentu

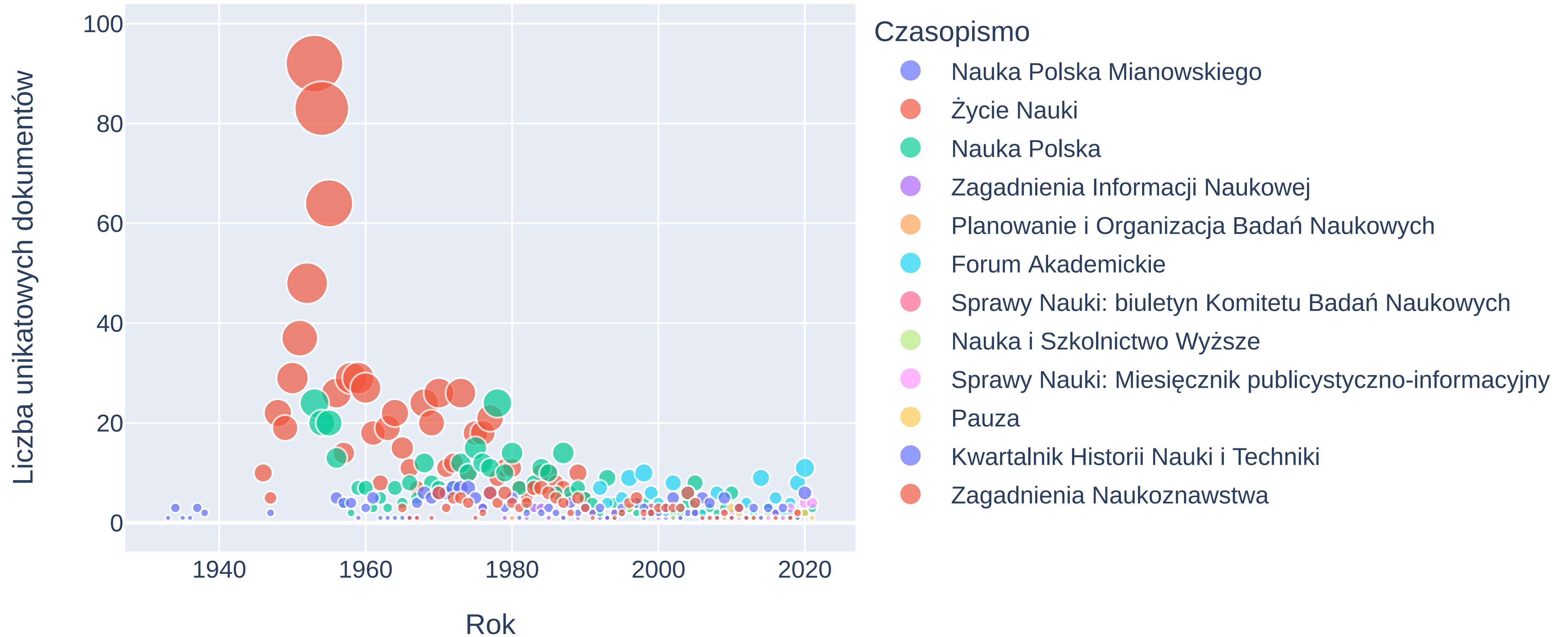


Podstawowe
narzędzia, które mogą
wspomóc różne nasze
badania

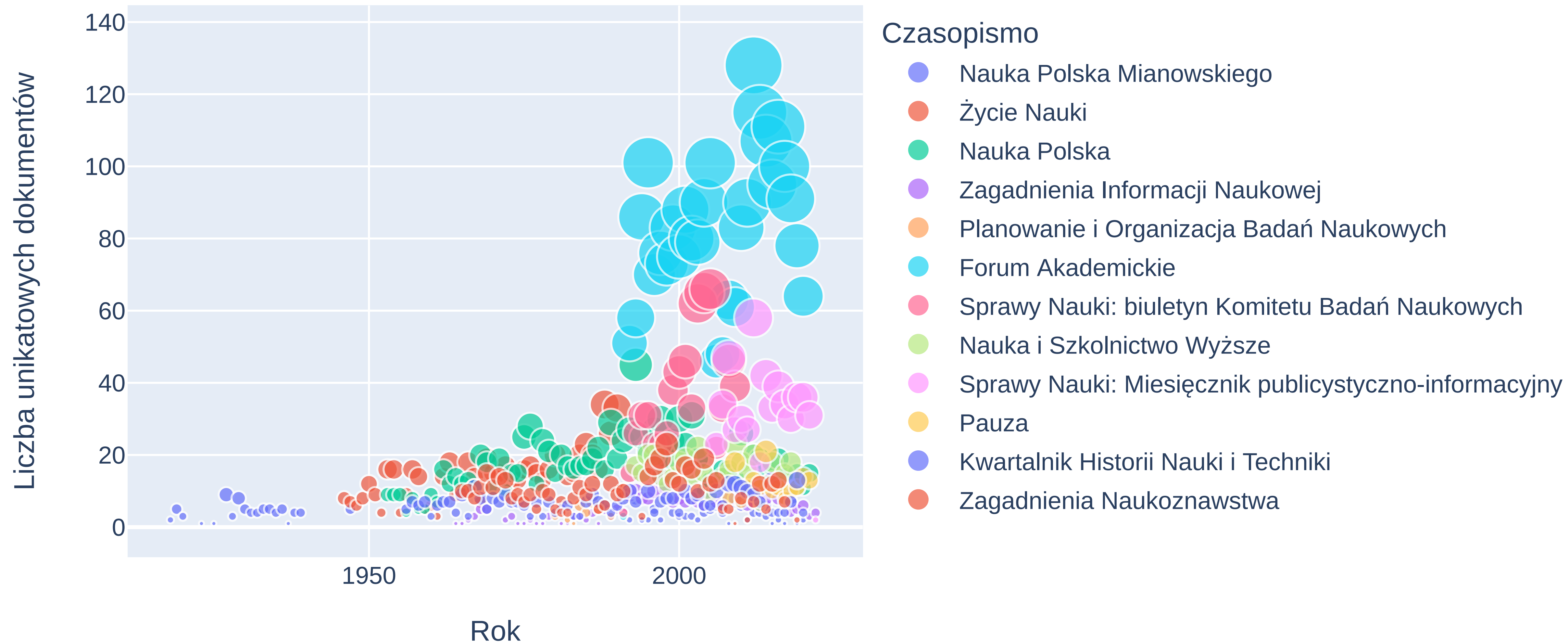
Wystąpienia słowa **ewaluacja** w zlematyzowanych dokumentach



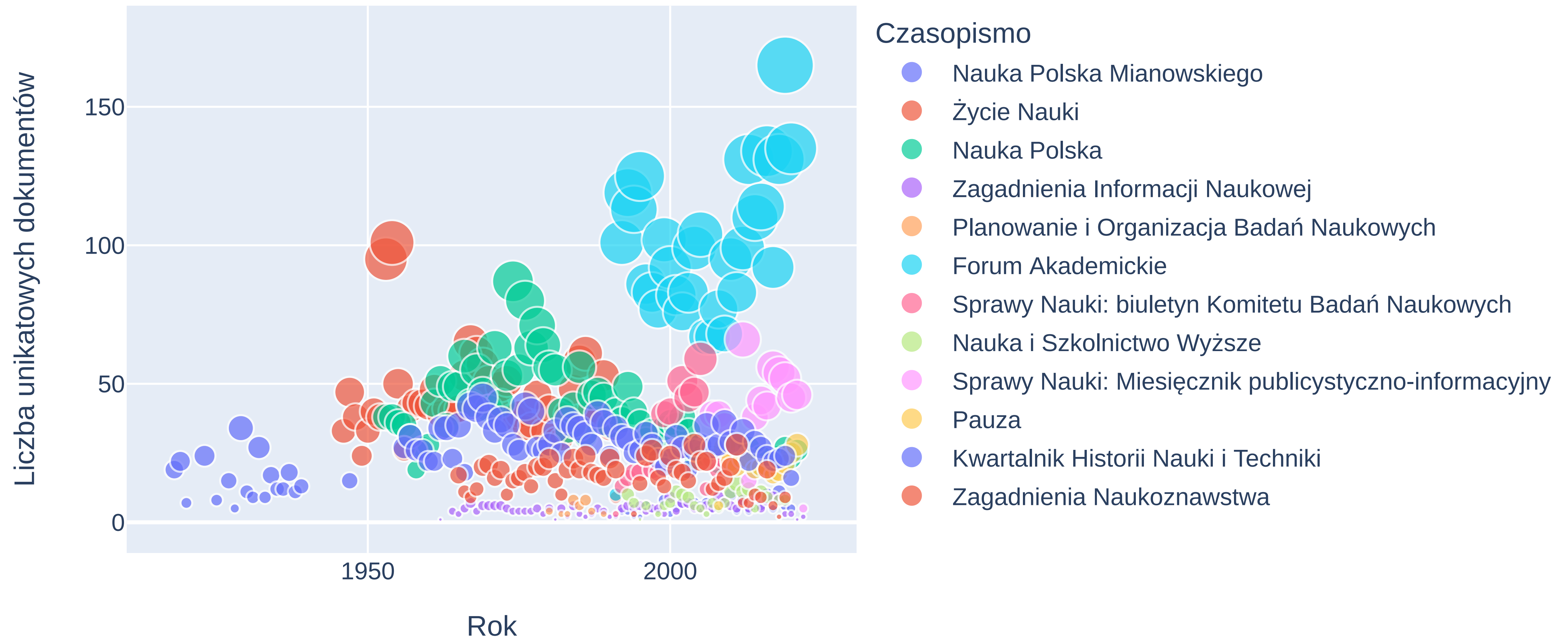
Wystąpienia słowa **marksizm** w zlematyzowanych dokumentach



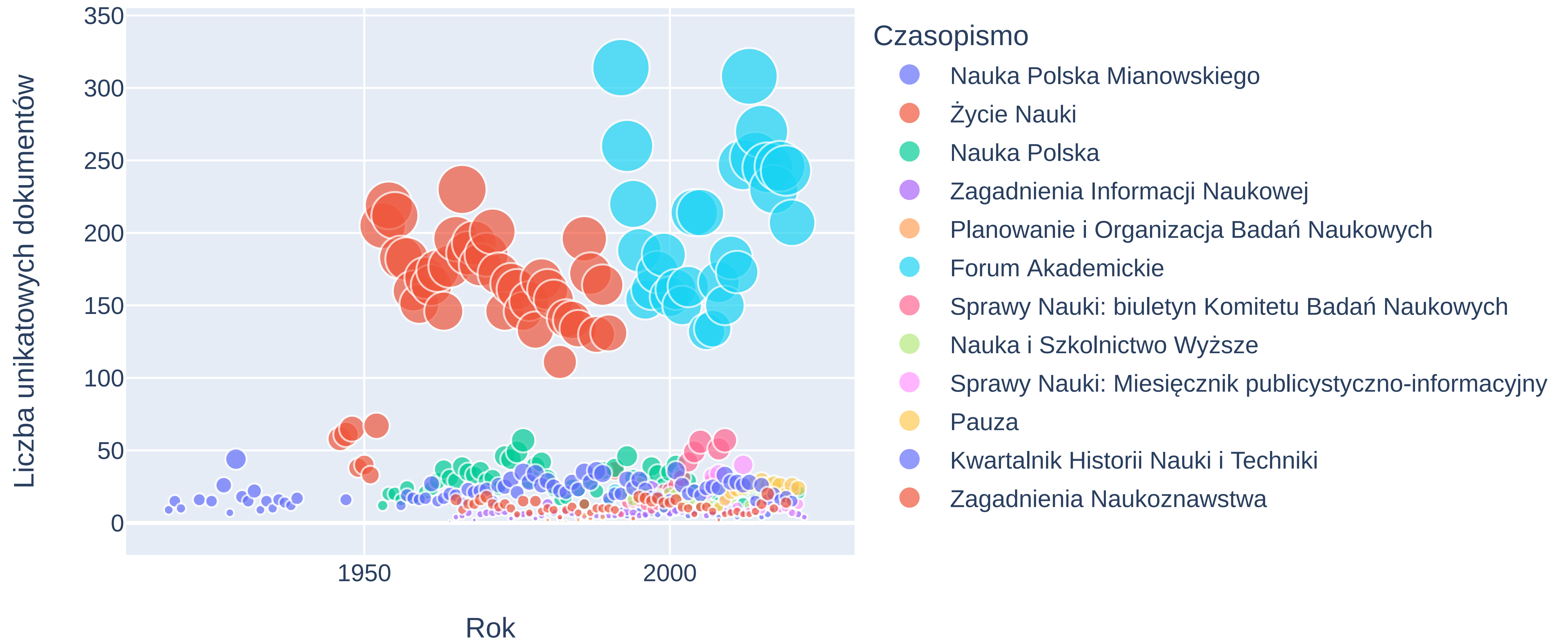
Wystąpienia słowa **rynek** w zlematyzowanych dokumentach



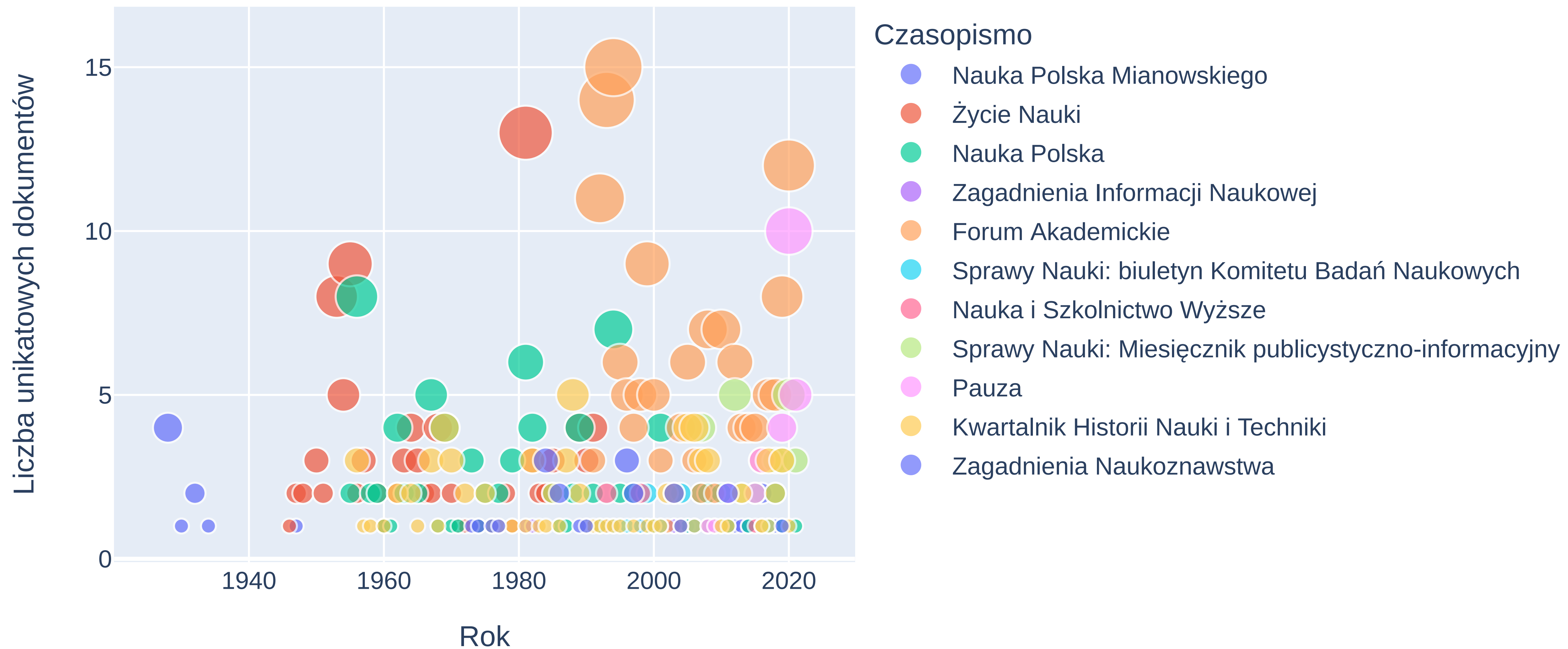
Wystąpienia słowa **wieś** w zlematyzowanych dokumentach



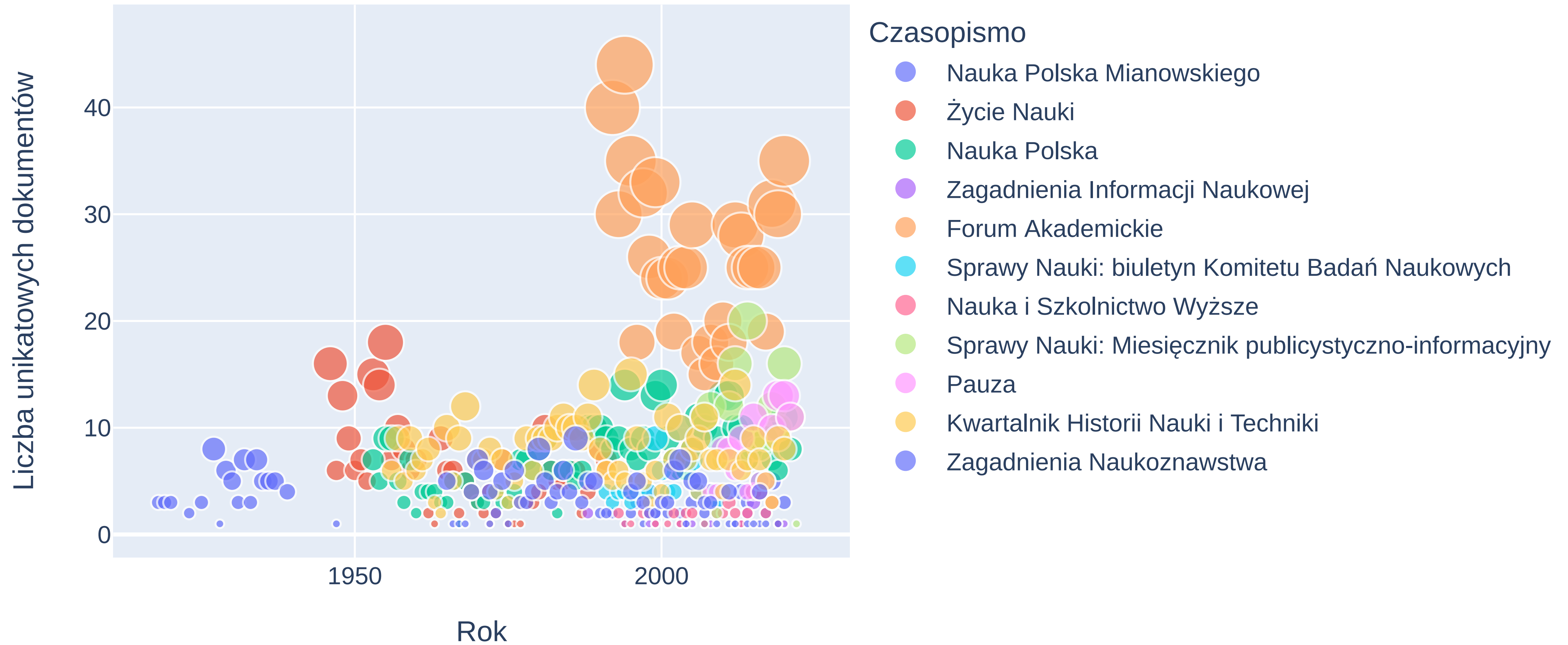
Wystąpienia słowa **student** w zlematyzowanych dokumentach



Wystąpienia słowa **strajk** w zlematyzowanych dokumentach



Wystąpienia słowa **protest** w zlematyzowanych dokumentach



Synonimy terminu **doskonałość** w obrębie *Zagadnień Naukoznawstwa* (Word2Vec model)

```
('doskonały', 0.56595778465271),  
( 'harmonia', 0.5603563785552979),  
( 'dostateczny', 0.5443614721298218),  
( 'gwarancja', 0.540084183216095),  
( 'utrzymanie', 0.5264930725097656),  
( 'wzniesienie', 0.5218774676322937),  
( 'pewność', 0.5203965902328491),  
( 'samodzielność', 0.5127161741256714),  
( 'osiągalny', 0.5111498832702637),  
( 'czystość', 0.5089118480682373)]
```

Weryfikowanie przypuszczeń

- ☑ Podobno fundacyjny dla polskiego naukoznawstwa tekst Floriana Znanieckiego z 1925 r. pt. *Przedmiot i zadania nauki o wiedzy* z „Nauki Polskiej” nie przyczynił się do wytworzenia dyskusji.
- ☑ Był nieczęsto cytowany (do tego dochodzą różnice w praktykach cytowań prawie 100 lat temu).
- ☑ Poszukajmy zatem „podobnych tekstów” (poprzez cosine similarity), abyśmy mogli je „ekspercko przeczytać”.

ID tekstu Znanieckiego



```
art = 'NPM_5_1_3'
coltext = 'Lemmagen'
ID_Column = 'ID Paper'
tresshold_sim = 0.5
ComparisonSimilarity(df_lem_art, coltext, ID_Column, tresshold_sim, art)
```

Cztery artykuły, dla których podobieństwo kosinusów wynosi ponad 0,5 (w skali od 0 do 1)



	ID Paper	Similarity coefficient
0	NPM_6_1_4	0.504107
1	NPM_11_1_5	0.557813
2	NPM_31_1_8	0.701021
3	NPM_49_1_5	0.506199

Podobny artykuł



And a similar article is:

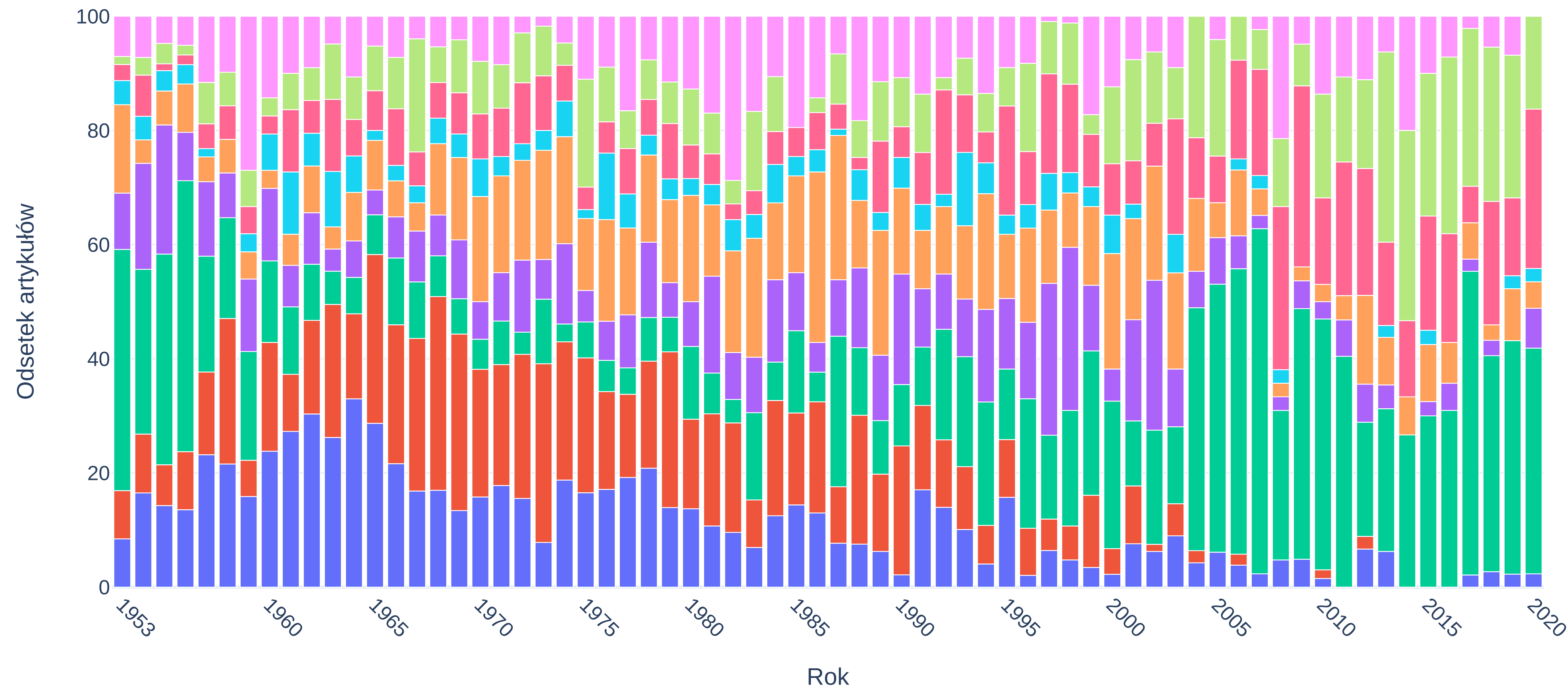
Unnamed: 0	ID Paper	Journal	Volume	Year	Title	Author
243	243 NPM_6_1_4	Nauka Polska Mianowskiego	1	1927	WYCHOWANIE PRACOWNIKA NAUKOWEGO	Zygmunt Mysłakowski

Analiza tematów

Topic modelling w oparciu
o Non-negative Matrix Factorization (NMF)

Nauka (Polska)
1953–2020

Rozkład tematów według lat



Tematy

- Temat 1: Badania techniczne i przyrodnicze
- Temat 2: Konferencje i sympozja naukowe
- Temat 3: Kultura, historia i literatura
- Temat 4: Struktury akademickie i organizacja nauki
- Temat 5: Rozwój gospodarczy i polityka ekonomiczna
- Temat 6: Działalność towarzystw naukowych
- Temat 7: Biologia, badania medyczne i zdrowie
- Temat 8: Uniwersytety, szkolnictwo wyższe i praca naukowa
- Temat 9: Produkcja rolno-spożywcza

Ewaluacja wyników modelowania (jak wyłoniłmy 9 tematów)

- ☑ analizowaliśmy relacje między tematami przy użyciu sieci grafów
- ☑ analizowaliśmy najważniejsze wyłonięte słowa dla danego tematu i tworzyliśmy ogólne etykiety dla tematów (tj. ich nazwy), aby dodatkowo zweryfikować potencjalne nachodzenie się tych tematów (w tym procesie każdy z autorów tworzył etykiety, a następnie uzgadnialiśmy wspólną etykietę lub wyjaśnialiśmy rozbieżności)
- ☑ sprawdzaliśmy, losowo wybrane, artykuły i sensowność ich zaklasyfikowania do poszczególnych tematów
- ☑ w przypadku analizy tematycznej „Nauki (Polskiej)” byliśmy w stanie posłużyć się porównaniem z Bibliografią Polskiej Naukometrii (BPN), która indeksuje blisko dwa tysiące publikacji polskich badaczy, sklasyfikowanych jako publikacje naukometryczne. W związku z tym wyłoniłmy z BPN wszystkie artykuły opublikowane w „Nauce (Polskiej)” i sprawdziliśmy, czy zostały przypisane do tematów, które można uznać za jako „naukometryczne”.

Perspektywy
rozwoju



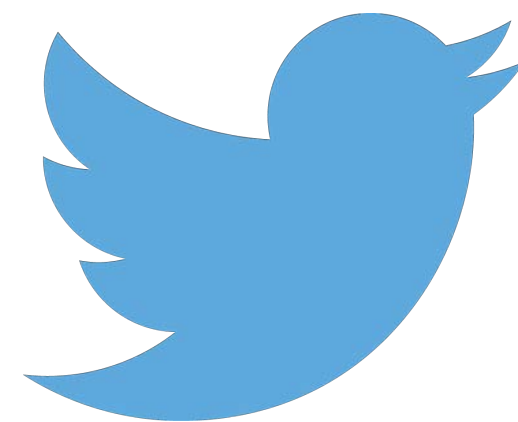
Co chcemy zrobić?

- ☑ Ze względu na prawa autorskie, nie może w pełni otworzyć korpusu. Ale jesteśmy w pełni otwarci na współpracę i wspólną pracę nad korpusem.
- ☑ Przede wszystkim warto rozbudować korpus o kolejne czasopisma, na łamach których rozwijały się dyskusje naukowawcze, takie jak „Studia Historiae Scientiarum” czy „Organon” (przy założeniu dodania innych języków).
- ☑ Jednym z większym wyzwań jest poprawienie jakości danych tekstowych po cyfrowym rozpoznaniu tekstu i w ten sposób ulepszenie procesu lematyzacji (w szczególności teksty z lat 20-tych XX wieku i z „powielaczy”).
- ☑ Stworzenie unikatowych identyfikatorów autorów pozwoli na przeprowadzenie dodatkowych analiz bibliometrycznych.
- ☑ Jednym z kierunków rozwoju może być dokonanie ekstrakcji cytowań z dokumentów, zarówno z bibliografii załącznikowej jak i przypisów dolnych.
- ☑ Nad korpusem pracuje obecnie doktorantka 1 roku w Szkole Doktorskiej Nauk Społecznych UAM, **mgr Sara Rotnicka**.

Dziękuję

emanuelkulczycki.com

Kontakt: emanuel@ekulczycki.pl



[@ekulczycki](https://twitter.com/ekulczycki)