INSTYTUT BADAŃ LITERACKICH POLSKIEJ AKADEMII NAUK
Institute of Literary Research Polish Academy of Sciences

CENTRUM HUMANISTYKI CYFROWEJ

CYFROWE LITERATUROZNAWSTWO

STARE PYTANIA – NOWE MOŻLIWOŚCI

DR MACIEJ MARYL
CENTRUM HUMANISTYKI CYFROWEJ IBL PAN

SPOTKANIA Z HUMANISTYKĄ CYFROWĄ
UNIWERSYTET WROCŁAWSKI
27.10.2017

www.ibl.waw.pl

ul. Nowy Świat 72, 00-330 Warsaw, Poland
phone/fax: (22) 826 99 45, (22) 65 72 895
e-mail: sekretariat@ibl.waw.pl

Kim są cyfrowi humaniści (w Polsce)?

- Europejska ankieta DARIAH przeprowadzona na przełomie 2014 i 2015
- Polska wersja opracowana przez CHC IBL PAN i PCSS
- 152 odpowiedzi z Polski
- Ankieta pokazuje generalne tendencje w środowisku osób zainteresowanych humanistyką cyfrową
  - Tłumaczenie: Maciej Maryl
  - Korekta: Piotr Wciślik (CHC), Michał Kozak, Marcin Werla, (PCSS)
  - Testy: Joanna Andrusiewicz, Łukasz Bukowiecki, Aleksandra Wójtowicz (CHC)


DARIAH VCC2 DiMPO

POLAND

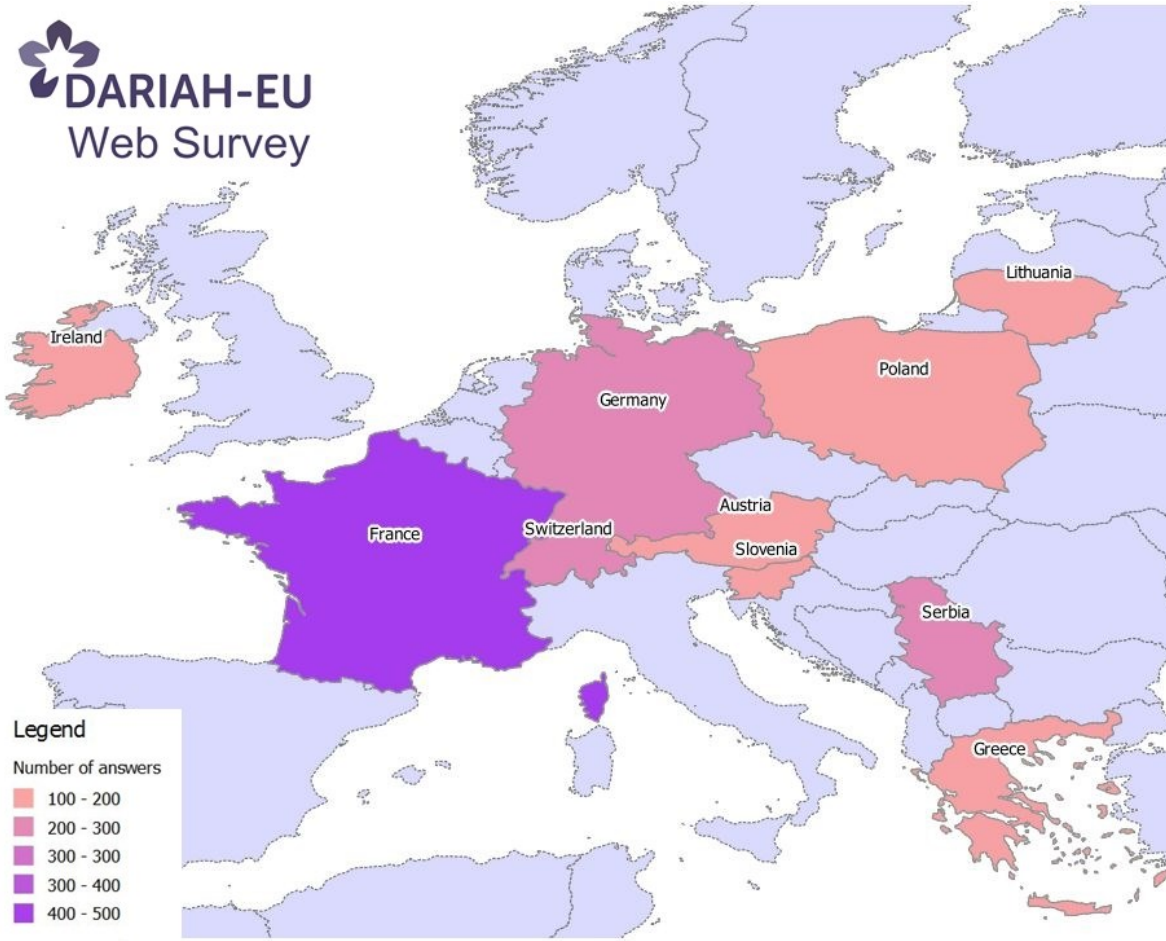
Report on the DARIAH Digital Practices in the Arts and Humanities Web Survey 2014.

Author: Maciej Maryl (Institute of Literary Research of the Polish Academy of Sciences)

Ankieta DARIAH

2,177 respondentów w Europie

10 języków

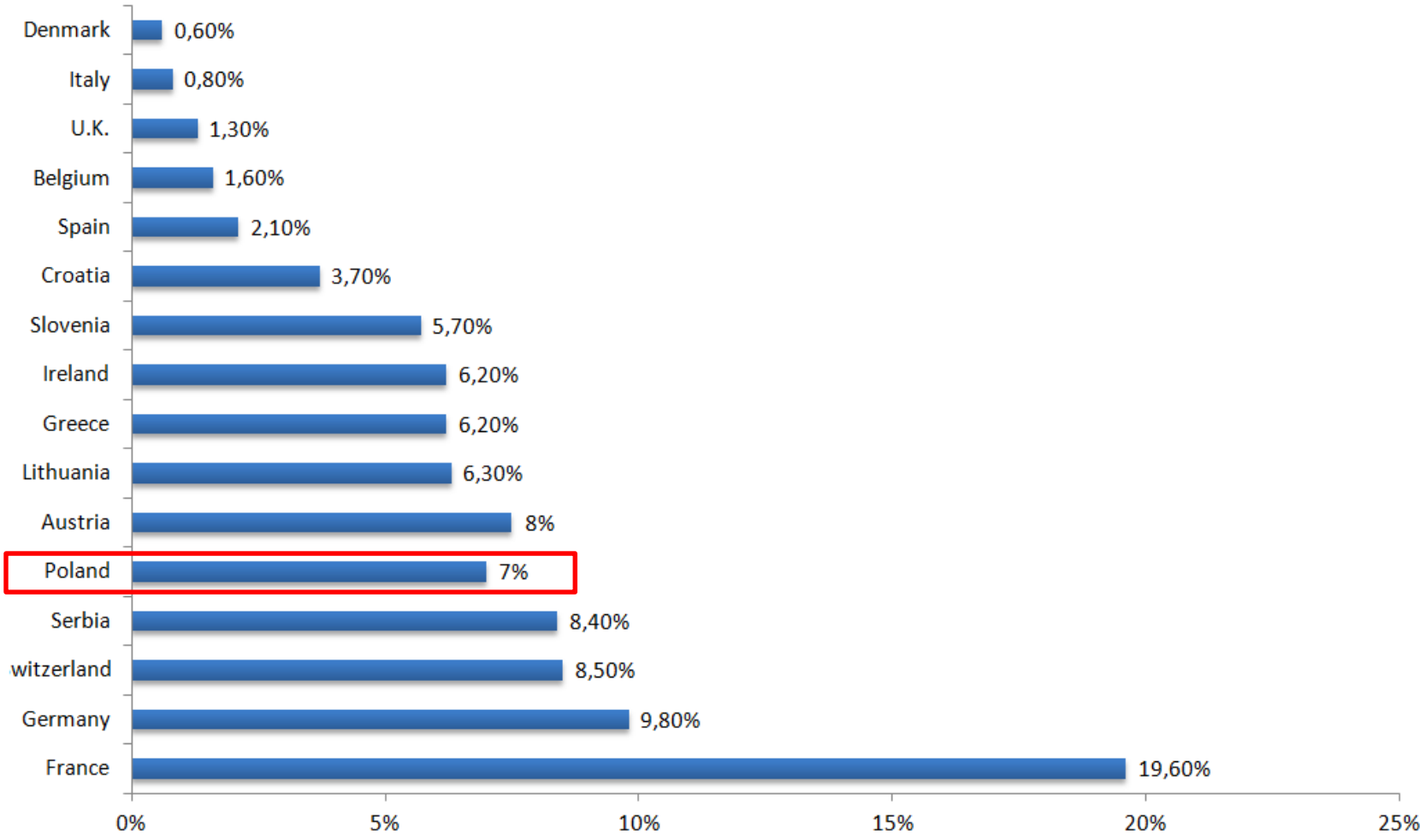


This work has been conducted as part of a larger survey, conducted in 2014-15 by the Digital Methods and Practices Observatory (DiMPO) Working Group of DARIAH-EU, the Digital Advanced Research Infrastructure for the Arts and Humanities in Europe project

Wyniki sondażu

- Główny raport ukaże się wkrótce
- Dallas, C., Chatzidiakou, N., Maryl, M., Bernardou, A., Clivaz, C., Cunningham, J., Dabek, M., et al. (2016). “Europejski Sondaż Praktyk Cyfrowych w Humanistyce i Naukach o Sztuce. Najważniejsze Wyniki (wersja Polska)”, przeł. Maciej Maryl. DOI:10.5281/zenodo.259522. [PDF]
- Maryl, M. (2017) “Kim są polscy humaniści cyfrowi?” *Teksty Drugie* 1 (163): 286-300. DOI: 10.18318/td.2017.1.24. [PDF]
- Maryl, M. (2014) „F5: odświeżanie filologii *Teksty Drugie* 2 (146): 9-20. [PDF]

Próba


Denmark 0,60%
Italy 0,80%
U.K. 1,30%
Belgium 1,60%
Spain 2,10%
Croatia 3,70%
Slovenia 5,70%
Ireland 6,20%
Greece 6,20%
Lithuania 6,30%
Austria 8%
Poland 7%
Serbia 8,40%
witzerland 8,50%
Germany 9,80%
France 19,60%
0%
5%
10%
15%
20%
25%

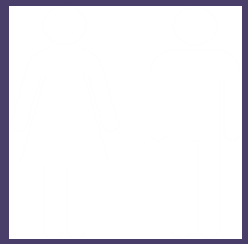
Dane demograficzne (próba polska i europejska)

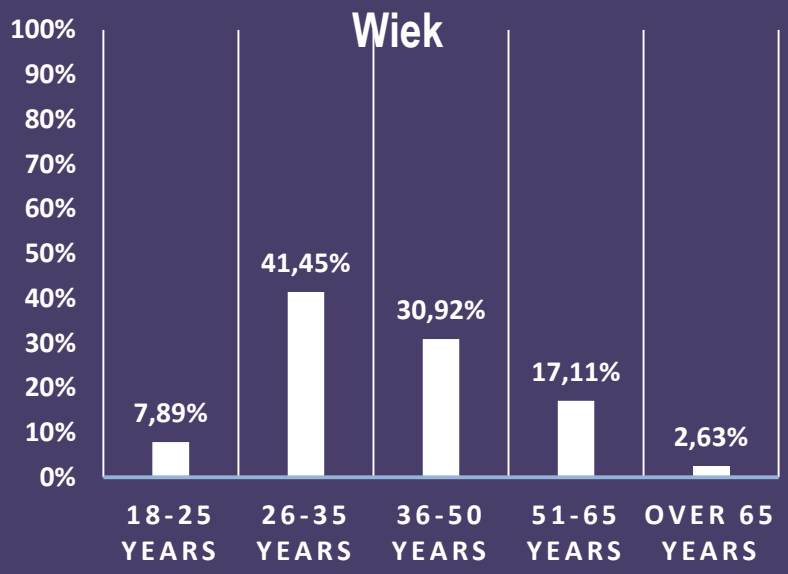
47,4 % kobiet

52,6 % mężczyzn

56,5% kobiet

43,5% mężczyzn

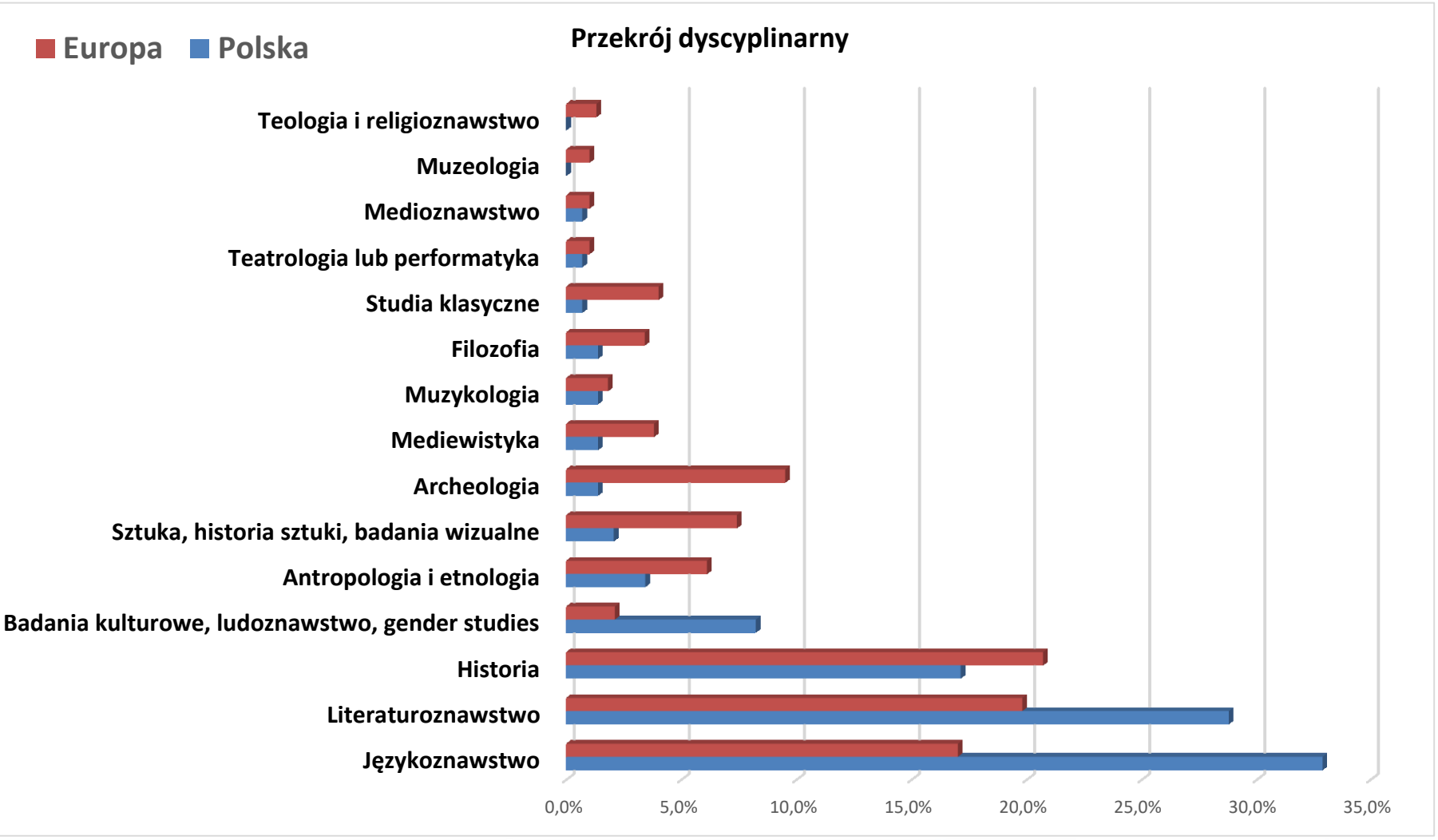




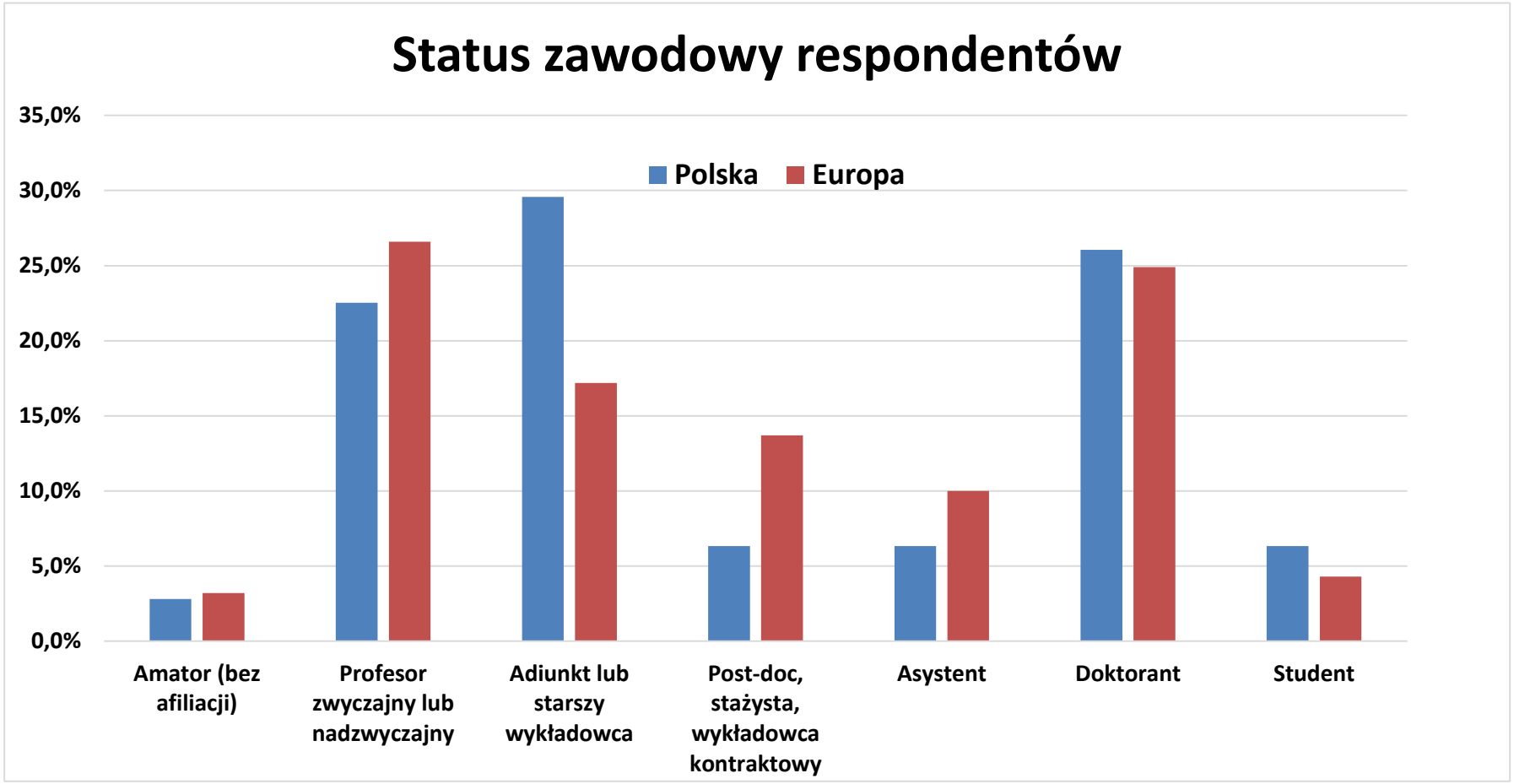
N=152

N=2169

Dyscypliny



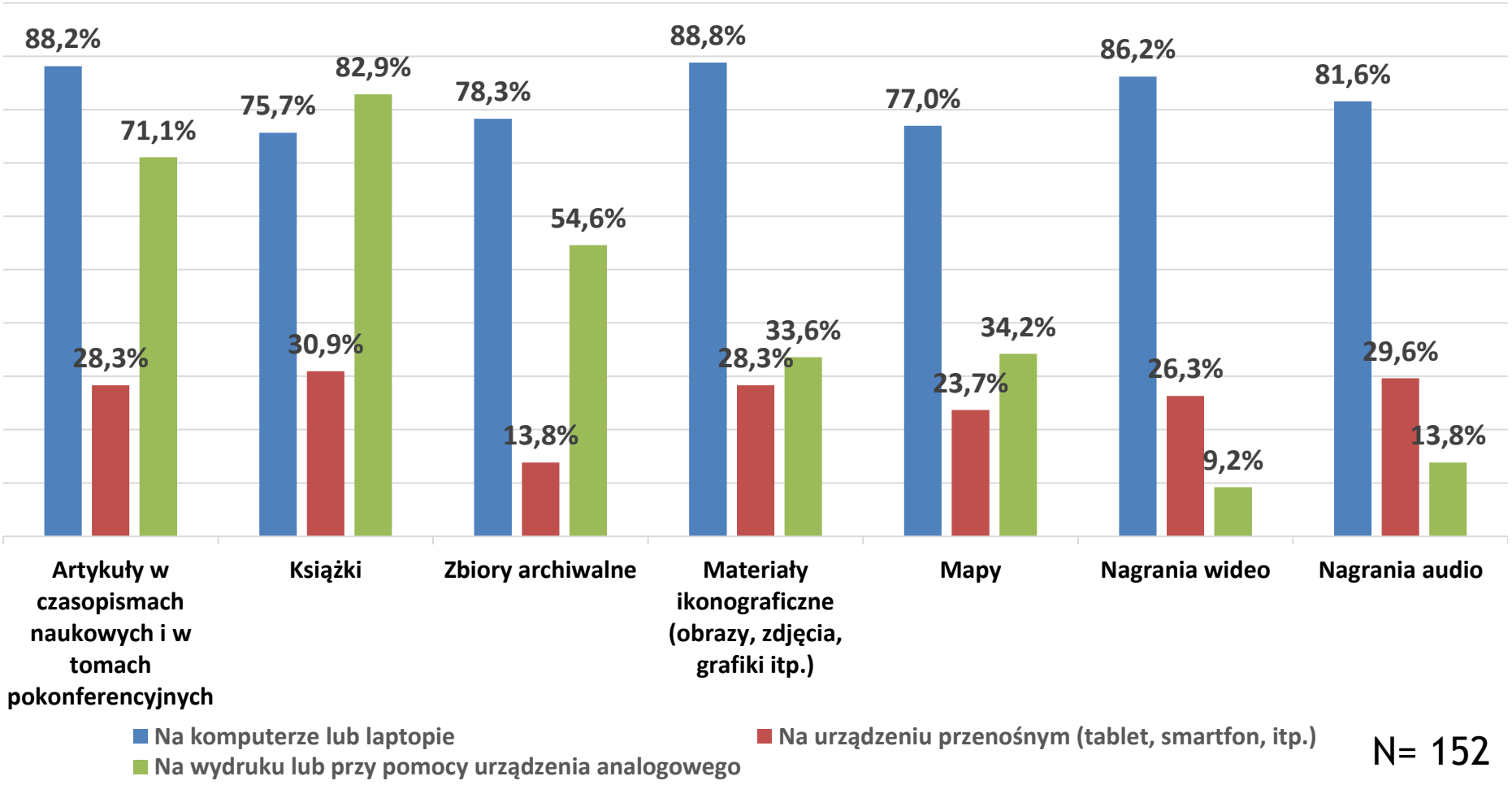
Status zawodowy



N= 142

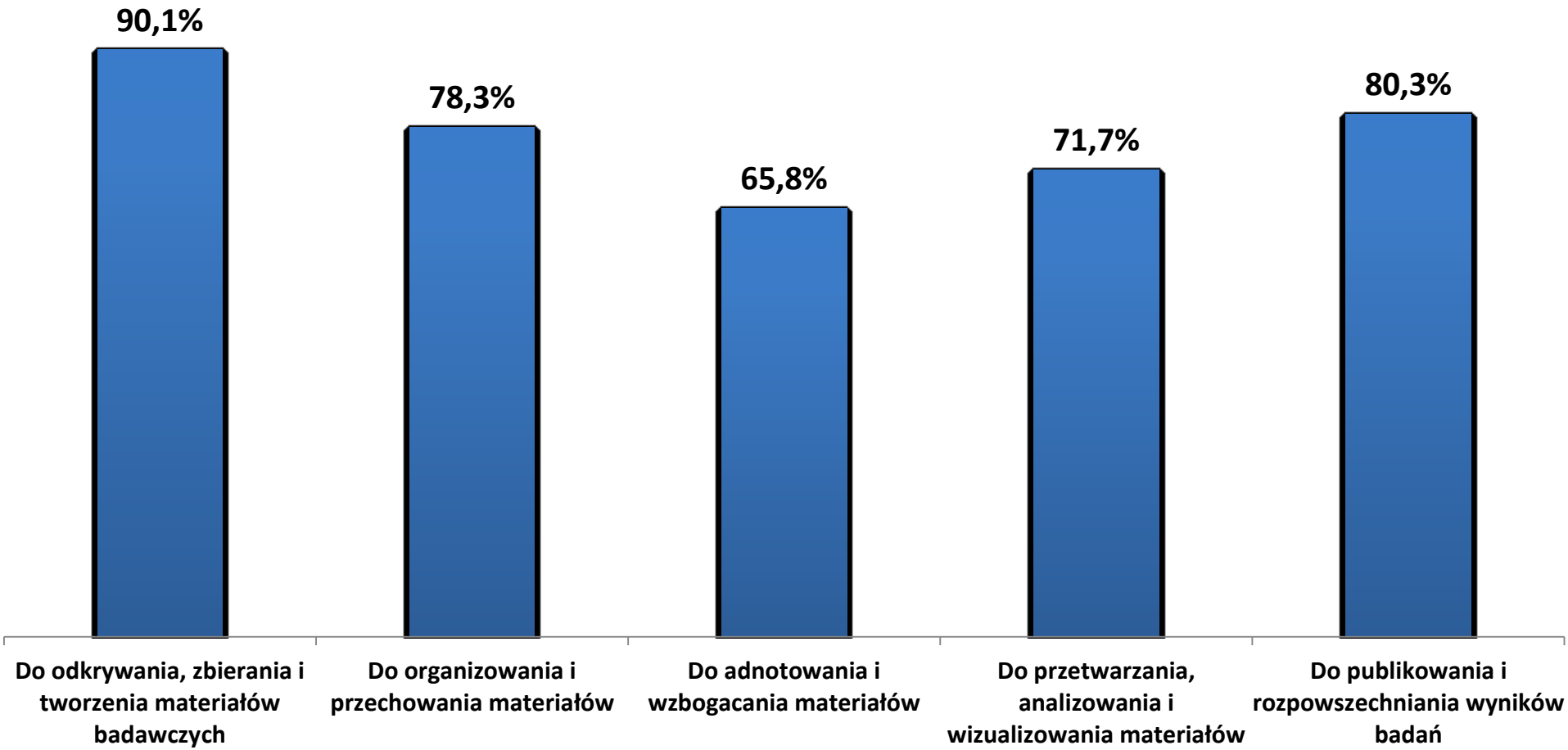
Korzystanie z mediów cyfrowych w pracy badawczej

W jaki sposób zapoznaje się Pan/i z następującymi materiałami?

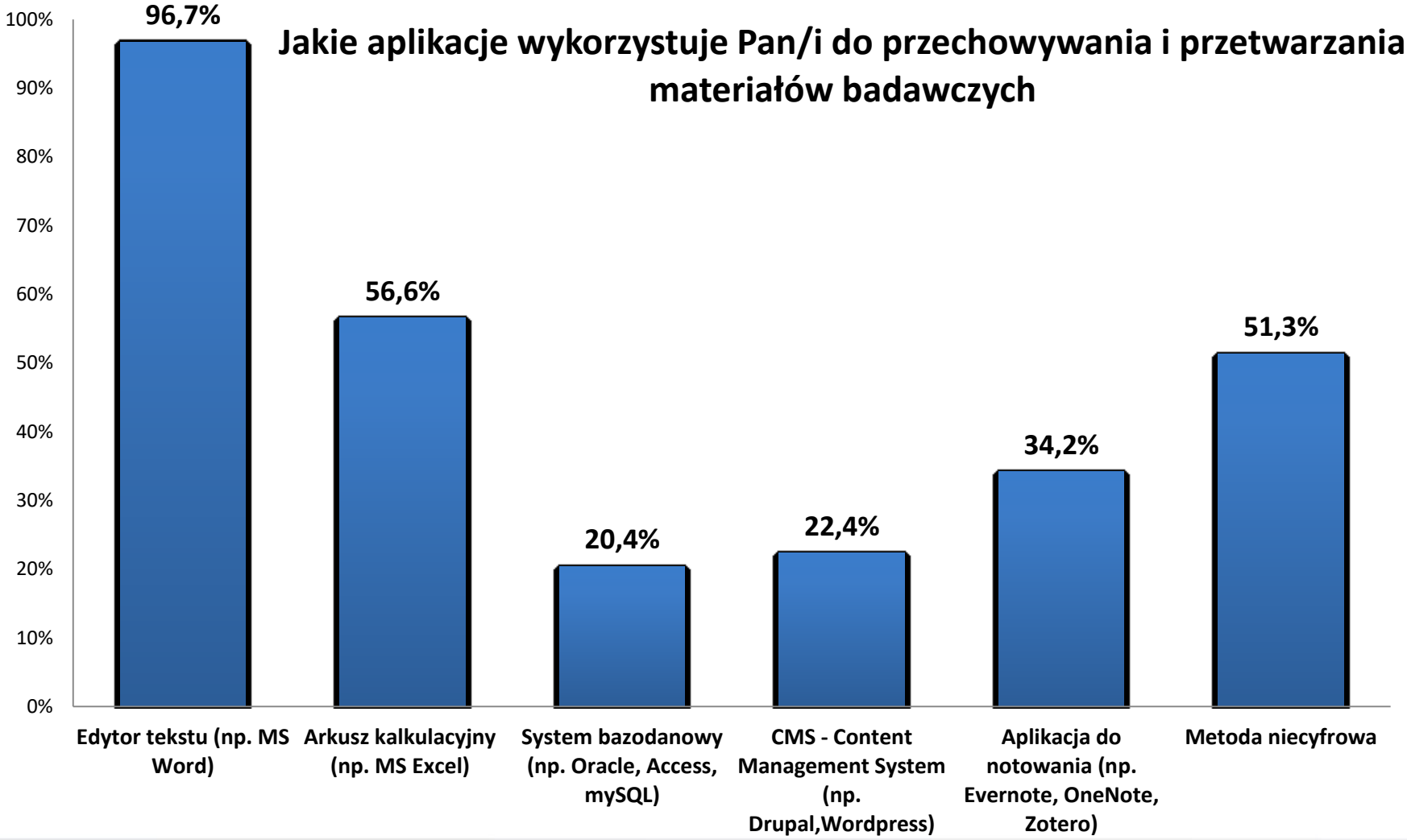


N= 152

Cele wykorzystania narzędzi cyfrowych w pracy badawczej

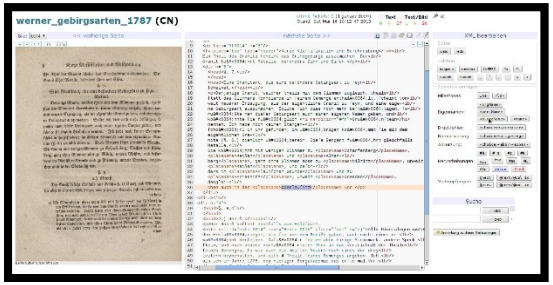


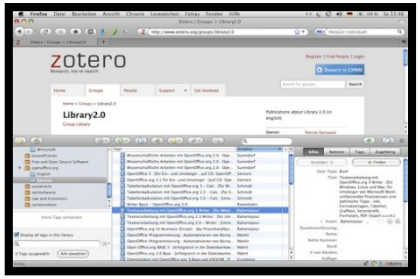
Aplikacje

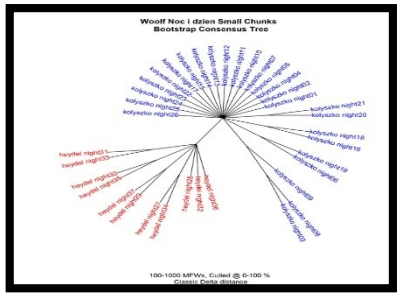


Metody i narzędzia cyfrowe

Od remediacji warsztatu po nowe metody i pytania badawcze







Co to jest humanistyka cyfrowa?

Wizualizacja wykonana przy pomocy wordle.net

Studia przypadku

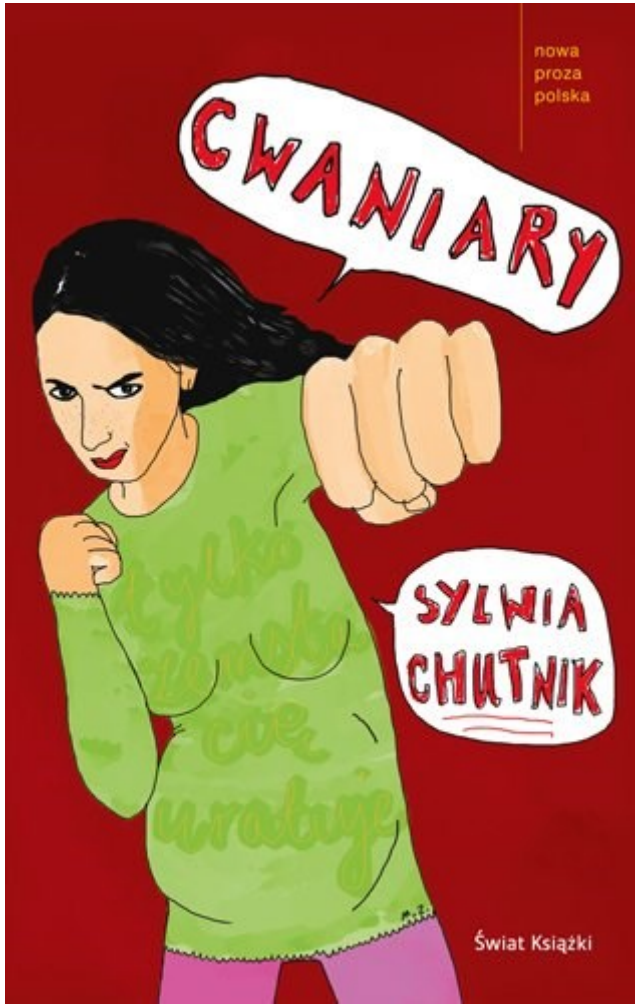
1. Porównanie wersji tekstu
2. Genologiczna analiza blogów
3. Analiza tekstów literaturoznawczych
4. Analiza metadanych

Studium przypadku 1
Porównanie wersji tekstu

- Maryl. M. (2016) „Startup literacki. Blog a powieść w odcinkach na przykładzie pierwowzoru Cwaniar Sylwii Chutnik” W: *Teksty kultury uczestnictwa.* Red. Dąbrówka, A., Maryl, M., Wójtowicz, A., Warszawa: Wydawnictwo IBL PAN, ss.85-110, doi: 10.18318/978-83-65573-14-8.5 [PDF]

Studium przypadku 1
Cwaniary Sylwii Chutnik

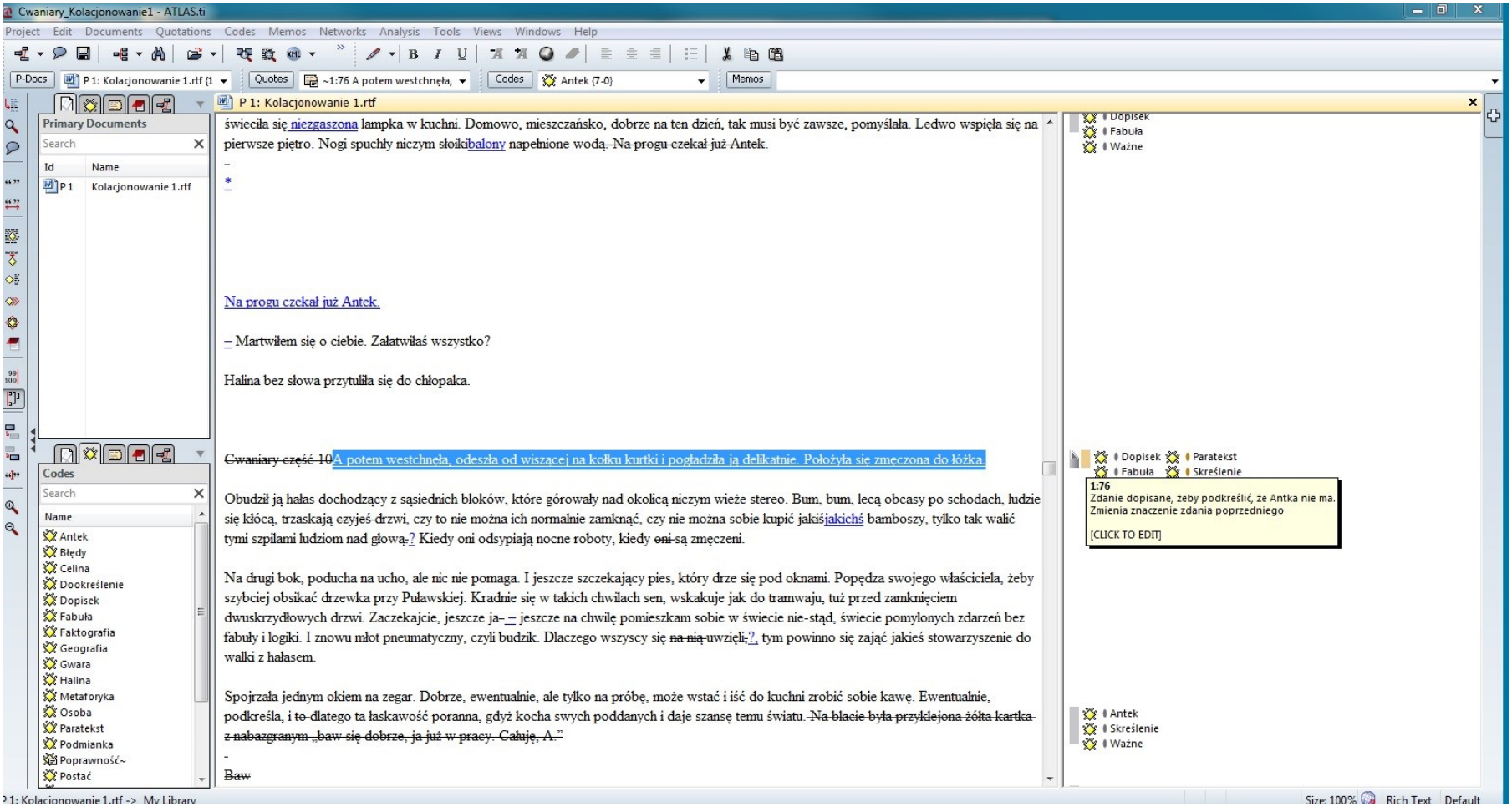
- *Cwaniary* (2012)
- Pierwowzór (pierwoDRUK?) blogowy
- 13 odcinków (czerwiec-sierpnień 2010)
- Pytanie badawcze: różnice między pierwowzorem blogowym a ostateczną wersją książki i ich wpływ na interpretację utworu.



Studium przypadku 1
Narzędzia i procedura badawcza

- Pobranie odcinków blogowych (WinHTTrack Website Copier 3.47-27)
- Pozyskanie wersji cyfrowej utworu (mobi) i jego konwersja na format tekstowy (rtf)
- Kolacjonowanie dwóch wersji utworu (Microsoft Word w pakiecie Microsoft Office Standard 2010 – komenda „porównaj dwie wersje dokumentu”)
- Wspomagana komputerowo analiza jakościowa (Atlas.ti 7.1.5).
- Analizy stylometryczne (R – stylo).

Studium przypadku 1
Analiza treści



Studium przypadku 1
Wnioski: różnice fabularne

Scena na cmentarzu

„Wtedy odkrywa się śmierć” (Antka)

1. Halina jedzie autobusem

2. Halina spotyka Celinę na Pradze

3. Razem z Celiną biją neonazistów

4. Piją w barze do rana

Historia Celiny

„Młode wdowy”

5. Spotykają Stefanię pobitą przez męża

6. Idą spać

7. Zabijają męża Stefanii (Biją dotkliwie)

„Ktoś tu kogoś spalił”

8. Szkoła rodzenia (Halina z Antkiem) (Halina z mamą)

9.Tatuaż (niedkończony)

Długie spotkanie w salonie kosmetycznym Stefy

10. Pojawia się Bronka

Studium przypadku 1
Wnioski: typy zmian

- **Językowe**
  - **Poprawki gramatyczne i stylistyczne**
  - ***Decorum* – bez neologizmów (patomatka), „podbrzusze” zamiast „jaj”**
  - **Zmiany semantyczne – „torby” zamiast „siatek”**
- **Faktograficzne**
  - **Nazwy ulic**
- **Redakcyjne?**
  - **Dookreślenie i rozwinięcie niejasnych fragmentów.**
  - **Wypełnianie luk w narracyjnych (leniwy czytelnik)**
- **Fabularne**
  - **Następstwo czasowe (wieczór vs. Ranek)**
  - **Antek**

Studium przypadku 1
Powrót Haliny do domu (blog)

Pożegnały się i Halina powoli wróciła do domu. Otwierając drzwi od klatki schodowej, spojrzała w stronę swojego mieszkania. W oknie świeciła się lampka w kuchni. Domowo, mieszczańsko, dobrze na ten dzień, tak musi być zawsze, pomyślała. Ledwo wspięła się na pierwsze piętro. Nogi spuchły niczym słoiki napełnione wodą. Na progu czekał już Antek.

– Martwiłem się o ciebie. Załatwiłaś wszystko?

Halina bez słowa przytuliła się do chłopaka.

Studium przypadku 1
Powrót Haliny do domu (książka)

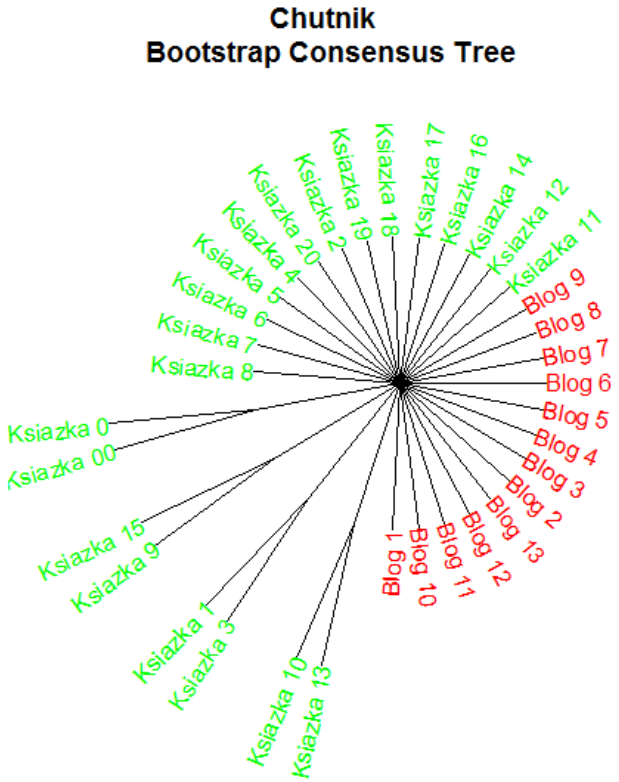
*Pożegnały się i Halina powoli wróciła do domu. Otwierając drzwi od klatki schodowej, spojrzała w stronę swojego mieszkania. W oknie świeciła się niezgaszona lampka w kuchni. Domowo, mieszczańsko, dobrze na ten dzień, tak musi być zawsze, pomyślała. Ledwo wspięła się na pierwsze piętro. Nogi spuchły niczym balony napełnione wodą. * Na progu czekał już Antek.*

– Martwiłem się o ciebie. Załatwiłaś wszystko?

Halina bez słowa przytuliła się do chłopaka.

A potem westchnęła, odeszła od wiszącej na kołku kurtki i pogładziła ją delikatnie. Położyła się zmęczona do łóżka.

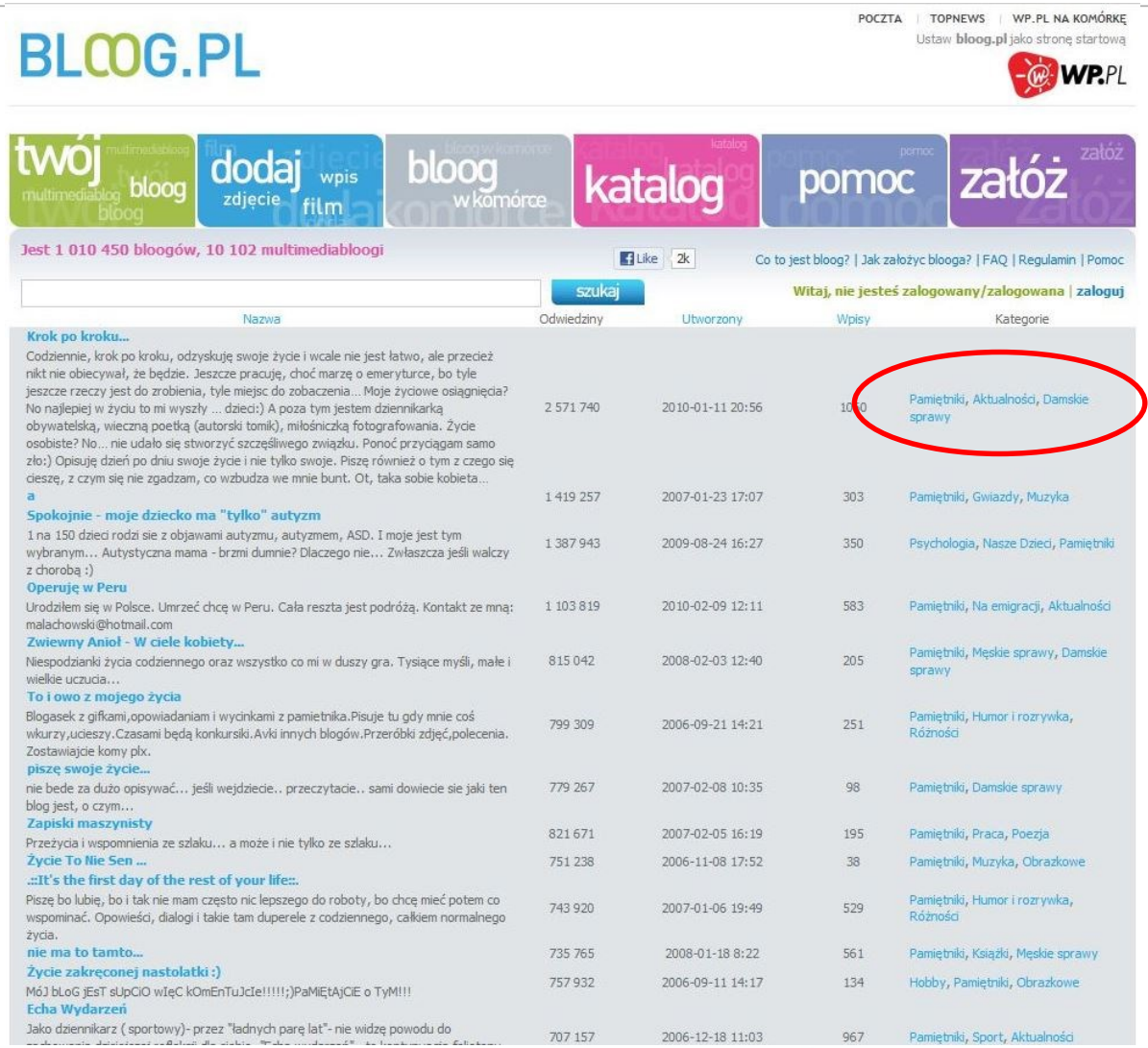
Studium przypadku 1
Analiza stylometryczna języka *Cwaniar*



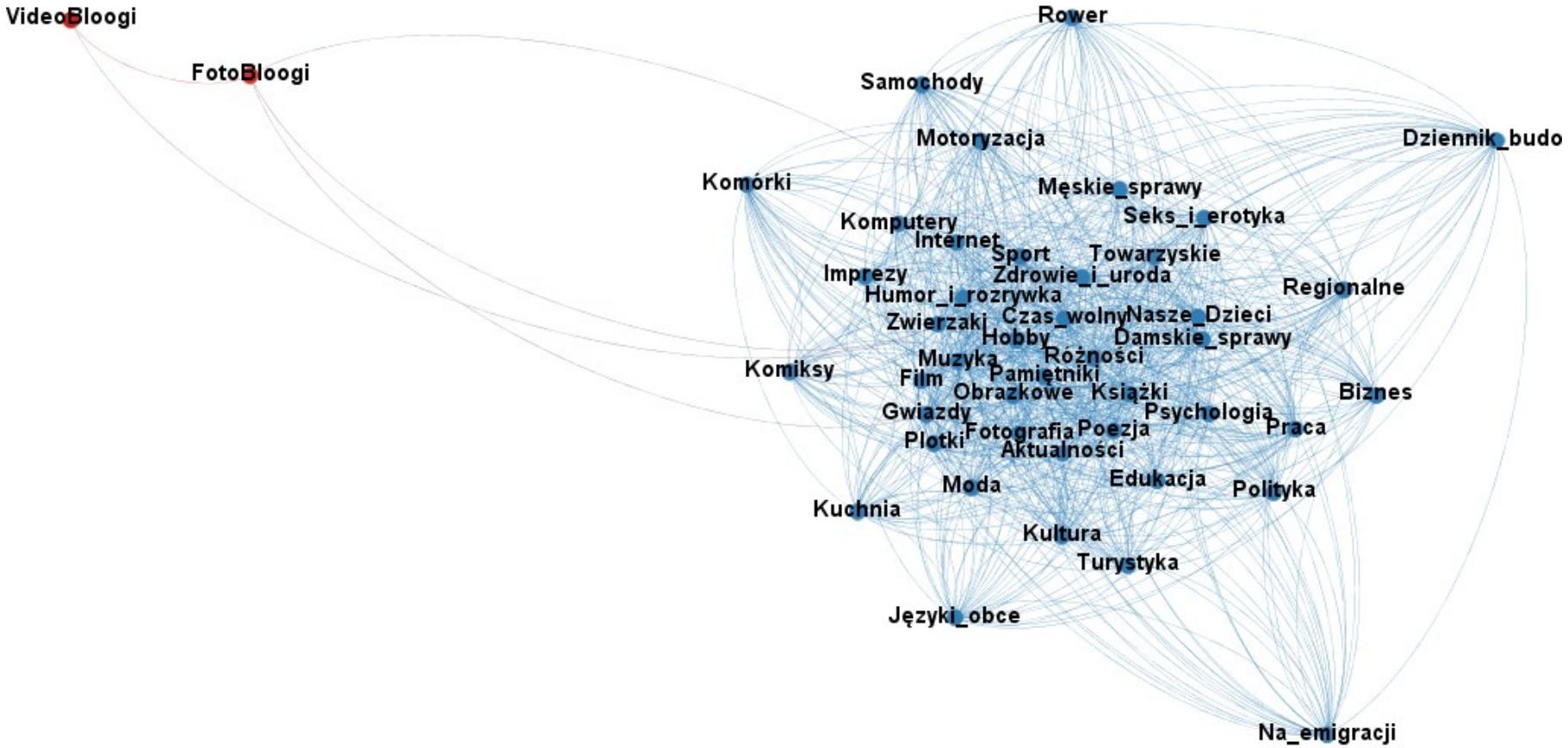
Studium przypadku 2
Analiza genologiczna blogów

- Współpraca: Krzysztof Niewiadomski, Maciej Kidawa, Maciej Piasecki, Ksenia Młynarczyk
- Maryl, M., Niewiadomski, K. i Kidawa, M. (2016). “Teksty elektroniczne w działaniu: typologia gatunków blogowych” *Zagadnienia Rodzajów Literackich* 59(118):2, ss. 51-72. [PDF]
- Maryl, M., Niewiadomski, K. and Kidawa, M. (2016). “Empirically Generated Typology of Weblog Genres”. *CLCWeb: Comparative Literature and Culture*, 18.2., June. [PDF]
- Maryl, M. (2016) „Tworzenie typologii gatunków piśmiennictwa multimedialnego na przykładzie blogów – propozycja metodologiczna” w: *Metody badań online*. Red. Piotr Siuda. Gdańsk: Wydawnictwo Naukowe Katedra, ss. 360-398. [PDF]
- Maryl, M., Piasecki, M., Młynarczyk, K. (2016) “Where Close and Distant Readings Meet: Text Clustering Methods in Literary Analysis of Weblog Genres.” W *Digital humanities 2016: Conference Abstracts*. Jagiellonian University & Pedagogical University, pp. 273-276. [PDF] [Online]
- Oleksy, M., Kocoń, J., Maryl, M., Piasecki, M. (2014) “Linguistic analysis of weblog genres” *Practical Applications of Language Corpora 2014: Book of Abstracts*, s. 79-81. [PDF]

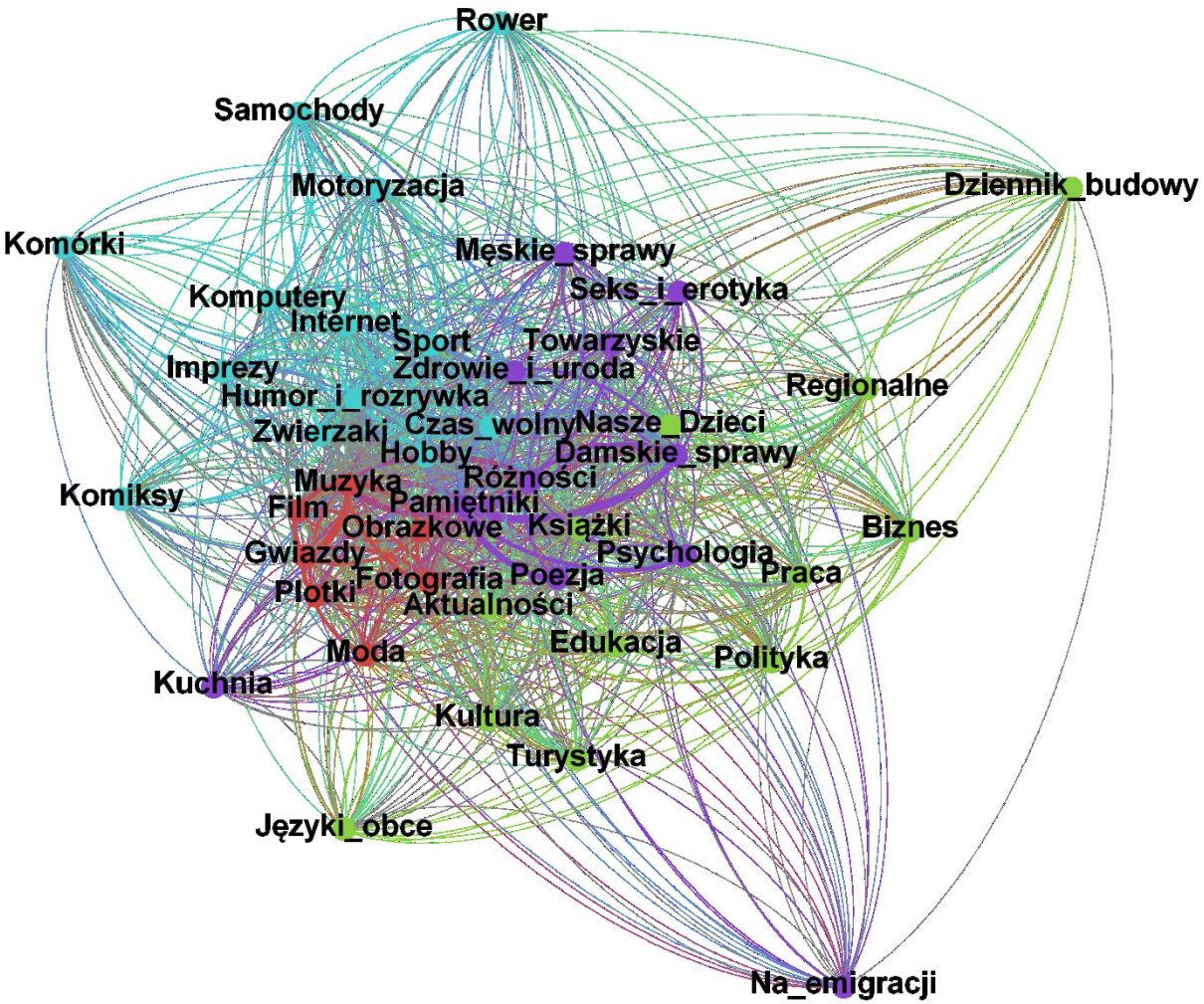
Analiza syntagmatyczna (N=88 252)



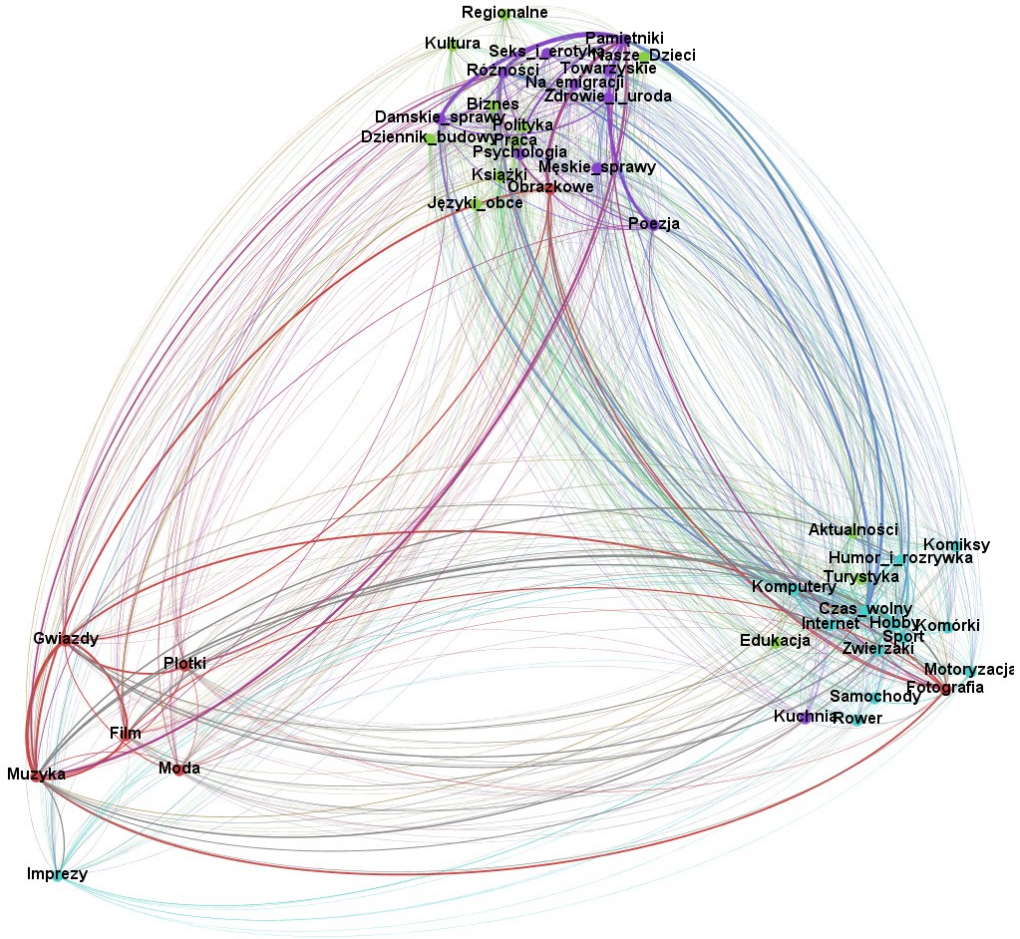
Ogólny obraz relacji między kategoriami (Force Atlas2)



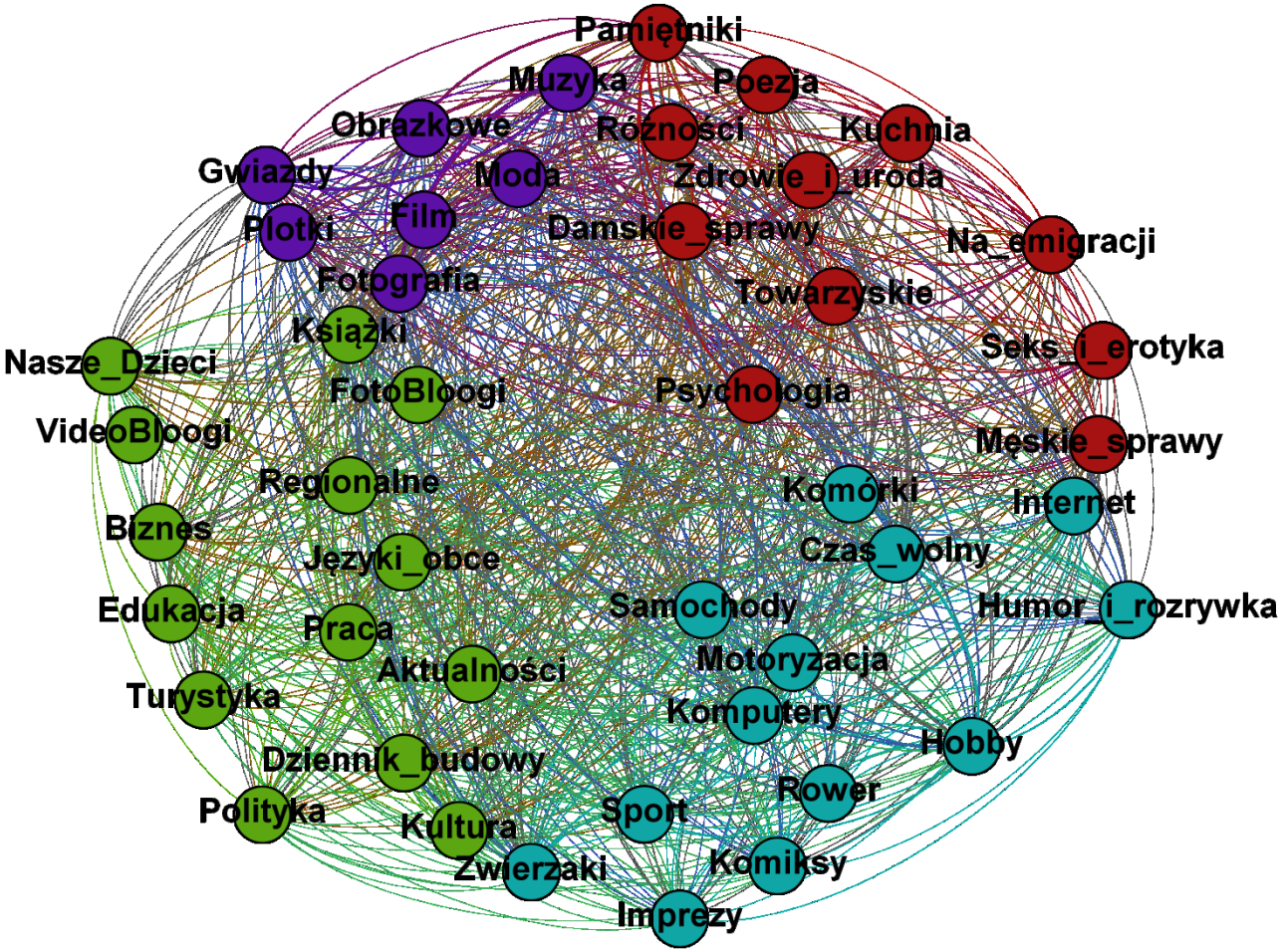
Widok modułów (Force Atlas2)



Skupiska (Open Ord)



Cztery typy (Radial Axis)



Analiza i interpretacja treści blogów
Procedura

1. Podział próby (322 blogi) na czteroosobowy zespół
2. Opracowanie wstępnej typologii
3. Uzgodnienie typologii
4. Przypisanie blogów do poszczególnych kategorii (trzech kodujących)
5. Uzgodnienie kategoryzacji

Gatunki blogowe

	Wystąpień	%Wszystkich	%Blogów
Krytyka	97	30,12%	31,80%
Porada	79	24,53%	25,90%
Diarystyka	55	17,08%	18,03%
Modelowanie	30	9,32%	9,84%
Refleksja	13	4,04%	4,26%
Informacja	13	4,04%	4,26%
Filtr	13	4,04%	4,26%
Fikcjonalność	5	1,55%	1,64%
Nieblog	15	4,66%	
Inne	2	0,62%	

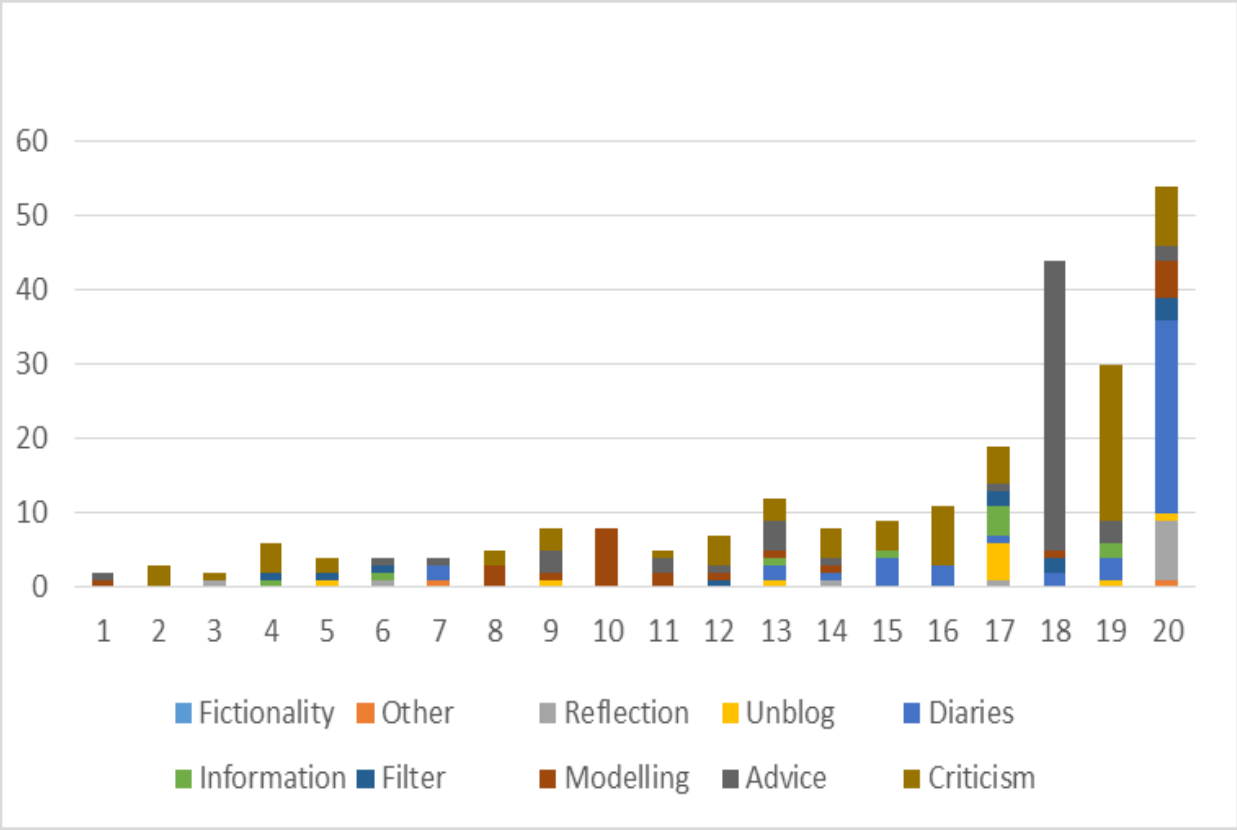
Empiryczna weryfikacja: analiza oparta na grupowaniu

- **Grupowanie**
  - **Pakiet *Cluto*** (Zhao & Karypis 2005)
    - **Grupowanie danych tekstowych**
  - **Pakiet *Stylo*** (Eder et al. 2013)
    - **wizualizacja grup (klastrów)**
- **Eksperymenty**
  - **poziom leksykalny: lematy i interpunkcja**
  - **poziom leksykalno-syntaktyczny: klasy gramatyczne i leksykalne**
  - **wyodrębnienie istotnych cech dla poszczególnych typów blogów**

Przegląd wyników

- Dość wysoka czystość (purity) klasterów
  - zakres: 58%-66%
    - tzn. ok. 60% blogów w danym klasterze należy do tego samego typu
- Entropia
  - zakres: 0.438-0.481,
    - tzn. mamy typy dominujące w klasterach ale inne blogi są rorzucone po różnych klasterach
- Brak dobrego dopasowania dla typów wyróżnionych drogą interpretacji

Analiza leksykalna



Cechy:

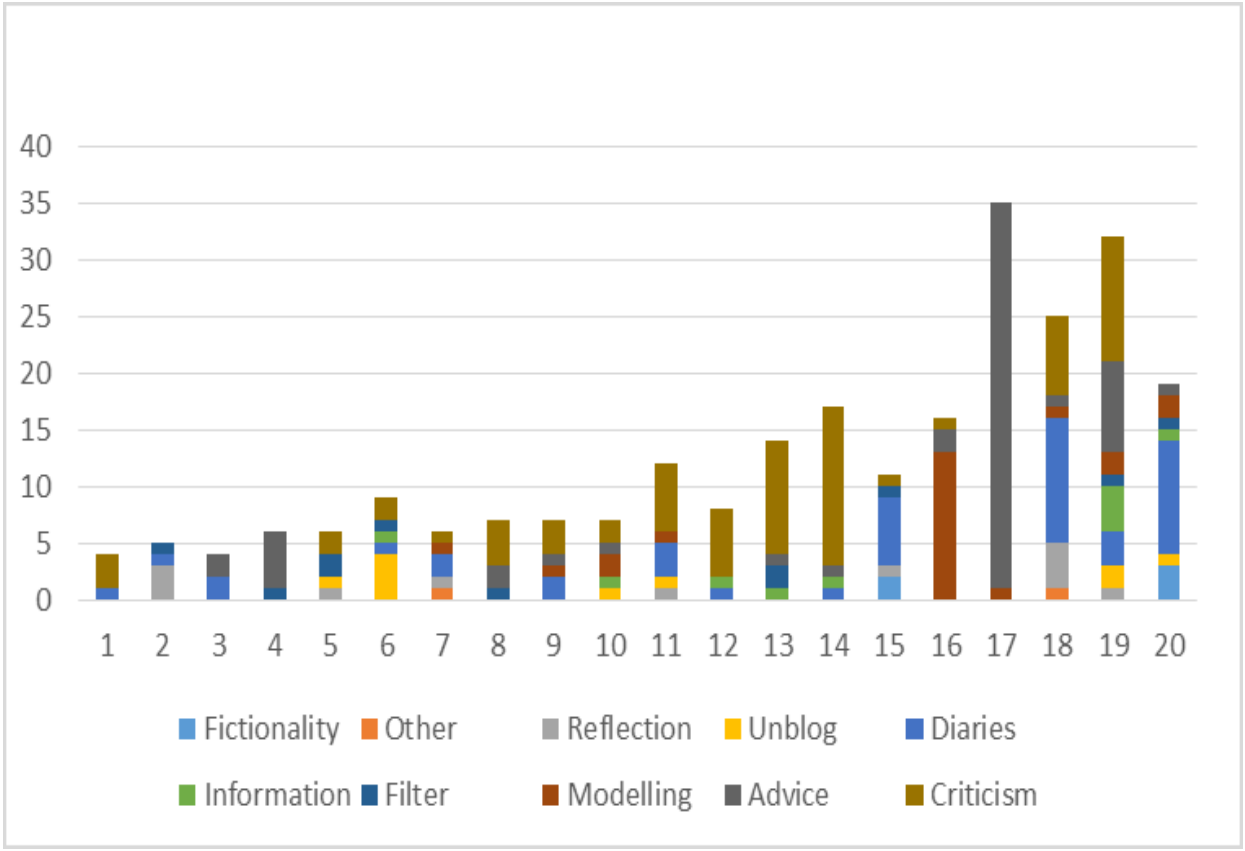
- 212 najczęstszych lematów,
- interpunkcja

Przetwarzanie:

- PMI weighting,
- ration similarity,
- graph clustering algorithm

Czystość: 58%
Entropia: 0.467

Analiza leksykalno-syntaktyczna



Cechy

- cechy leksykalne
- klasy gramatyczne
- trigramy

Przetwarzanie:

- PMI weighting,
- ration similarity,
- graph clustering algorithm

Czystość: 60.4%
Entropia: 0.438

Cechy charakterystyczne

Genre	Linguistic features
Dzienniki	1. & 2. osoba, słownictwo związane z „ja”
Fikcjonalność	Czas przeszły, 3. osoba
Filtry	Interpunkcja, rzeczowniki
Informacja	Formy bezosobowe czasowników, 3. osoba
Krytyka	Słownictwo subiektywna („ja”, „moje”, wykorzystanie spójników wskazujących na wnioskowanie logiczne
Modelowanie	Wtrącenia (np. „eh”), wykrzykniki, 1. & 2. osoba, słownictwo: „mój”, „rzecz”, „nowy”, „dlaczego”, „dlatego”
Porada	Bezokoliczniki, imiesłowy, słownictwo wskazujące na konkretne działania („około”, „duży”, „mały”)
Refleksja	1.&2. osoba, „sobie”, „zawsze”, „wszystko”

Studium przypadku 3
Analiza tekstu

- Współpraca: Tomasz Walkowiak, Maciej Piasecki, Maciej Eder
- Maryl, M. (2016) „Tekstów świat. Przyczynek do makroanalitycznej monografii czasopisma literaturoznawczego" w: *Projekt na daleką metę. Prace ofiarowane Ryszardowi Nyczowi.* Red. Anna Nasiłowska i Zdzisław Łapiński, Warszawa: Wydawnictwo IBL, ss.443-462. [PDF]
- Piasecki, M., Walkowiak, T., Maryl, M., (2017) "Literary Exploration Machine. New Tool for Distant Readers of Polish Literature." W *Digital humanities 2017: Conference Abstracts*. McGill University. [PDF]
- Maryl, M., Eder, M. (2017) "Topic Patterns in an Academic Literary Journal: The Case Of *Teksty Drugie*" W *Digital humanities 2017: Conference Abstracts*. McGill University. [PDF]

Literacki Eksplorator Maszynowy (LEM)

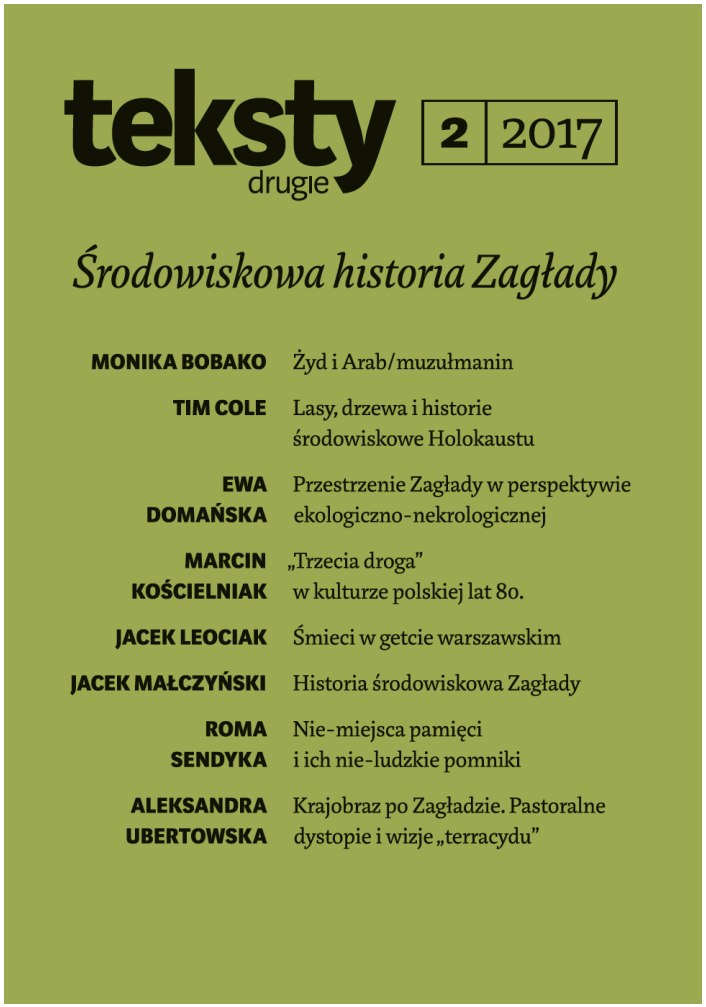
- agregator istniejących narzędzi dla języka polskiego
- pozwala na wykorzystanie narzędzi stworzonych dla innych języków (CLUTO, Mallet)
- proste procedury nie wymagające umiejętności programistycznych (załaduj korpus, wybierz parametry, odbierz wynik)
- nastawienie na użytkownika – LEM jest rozwijany przez studia przypadku poświęcone konkretnym problemom badawczym
- LINK: http://ws.clarin-pl.eu/lem.shtml?en

http://ws.clarin-pl.eu/lem.shtml?en

LEM: Funkcje

- Przekształcanie tekstu
  - Lematyzacja
  - Oznaczanie części mowy
- Statystyki morfosyntaktyczne
  - Charakterystyka czasowników (osoba i rodzaj)
  - Frekwencja lematów i części mowy
- Semantyka
  - Nazwy własne z frekwencjami
  - Ujednoznacznienie słów (Wordnet)
- Analiza stylometryczna: *WebSty*

Case Study: Teksty Drugie (*Second Texts*)



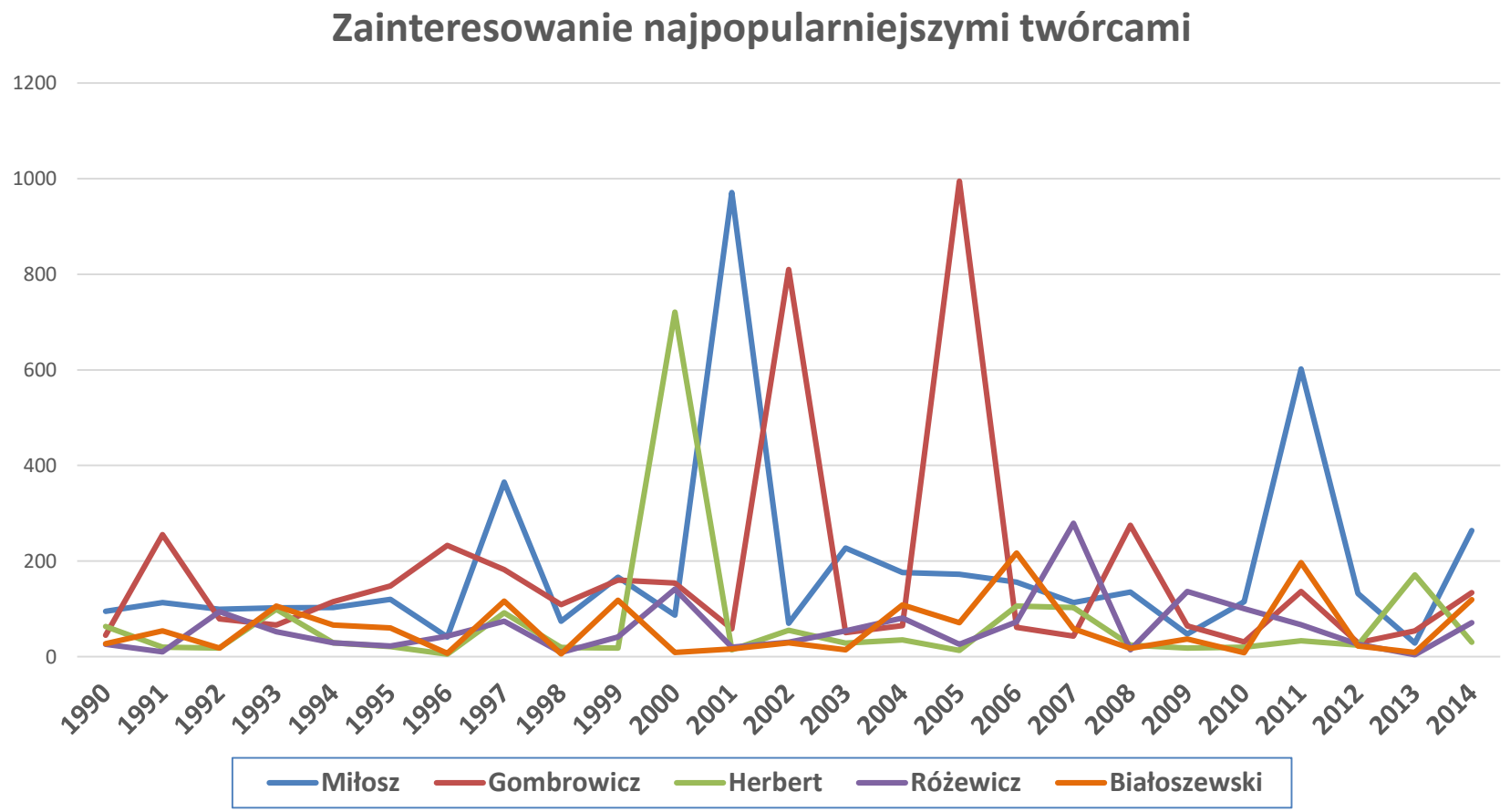
- Dwumiesięcznik
- Od 1990
- 262 numery i 2804 artykuły (do końca 2015)
- 1237 twórców (autorzy, tłumacze)
- Kluczowe zagadnienia badań literackich i kulturowych
- www.tekstydrugie.pl

Procedura

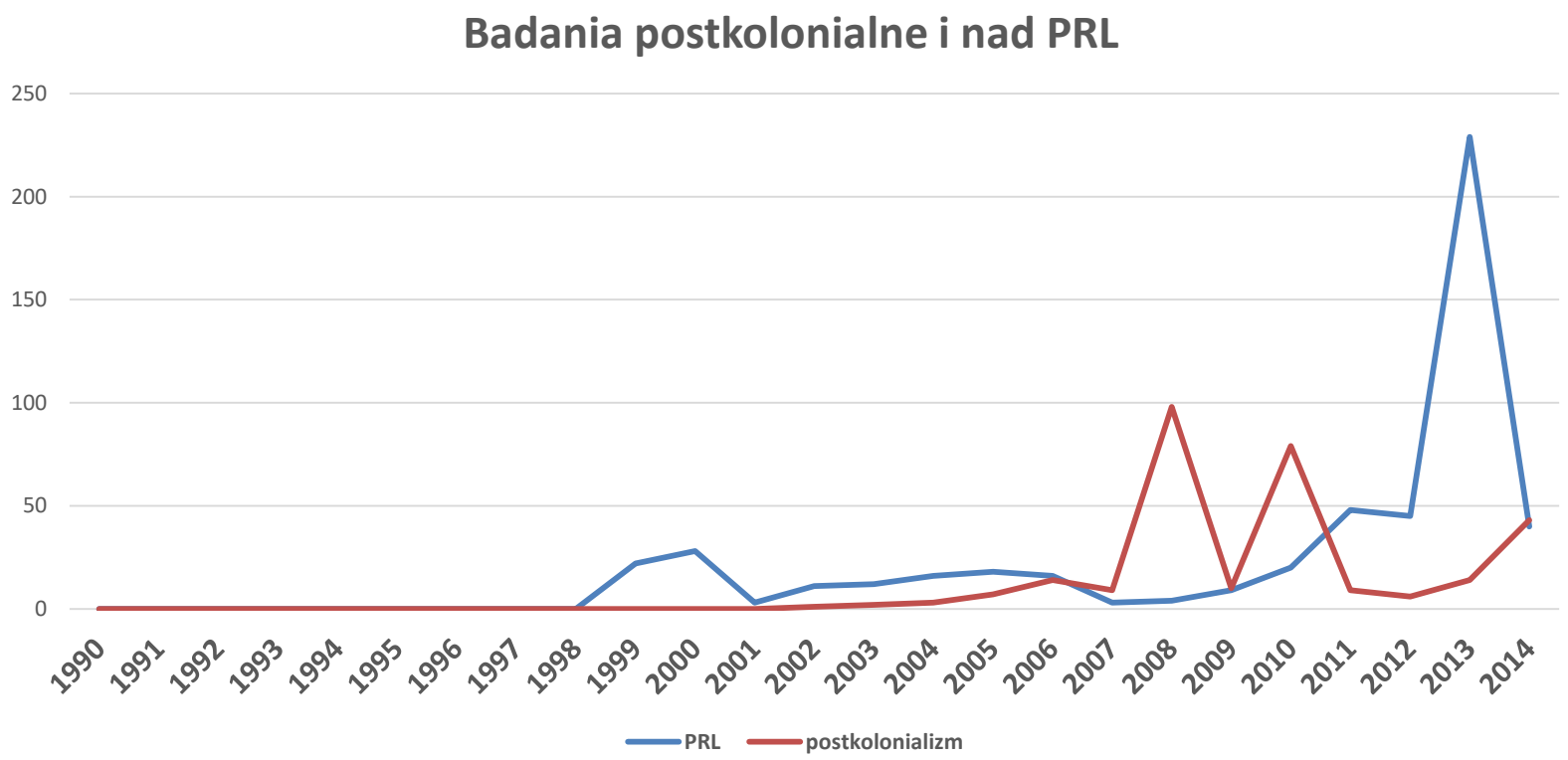
- Faza 1. OCR i czyszczenie korpusu
- Faza 2. Przetwarzanie i ręczna analiza frekwencji
  - Lematyzacja
  - Analiza wybranych lematów
- Faza 3. Eksploracja
  - Topic modelling

Proste chmury słów

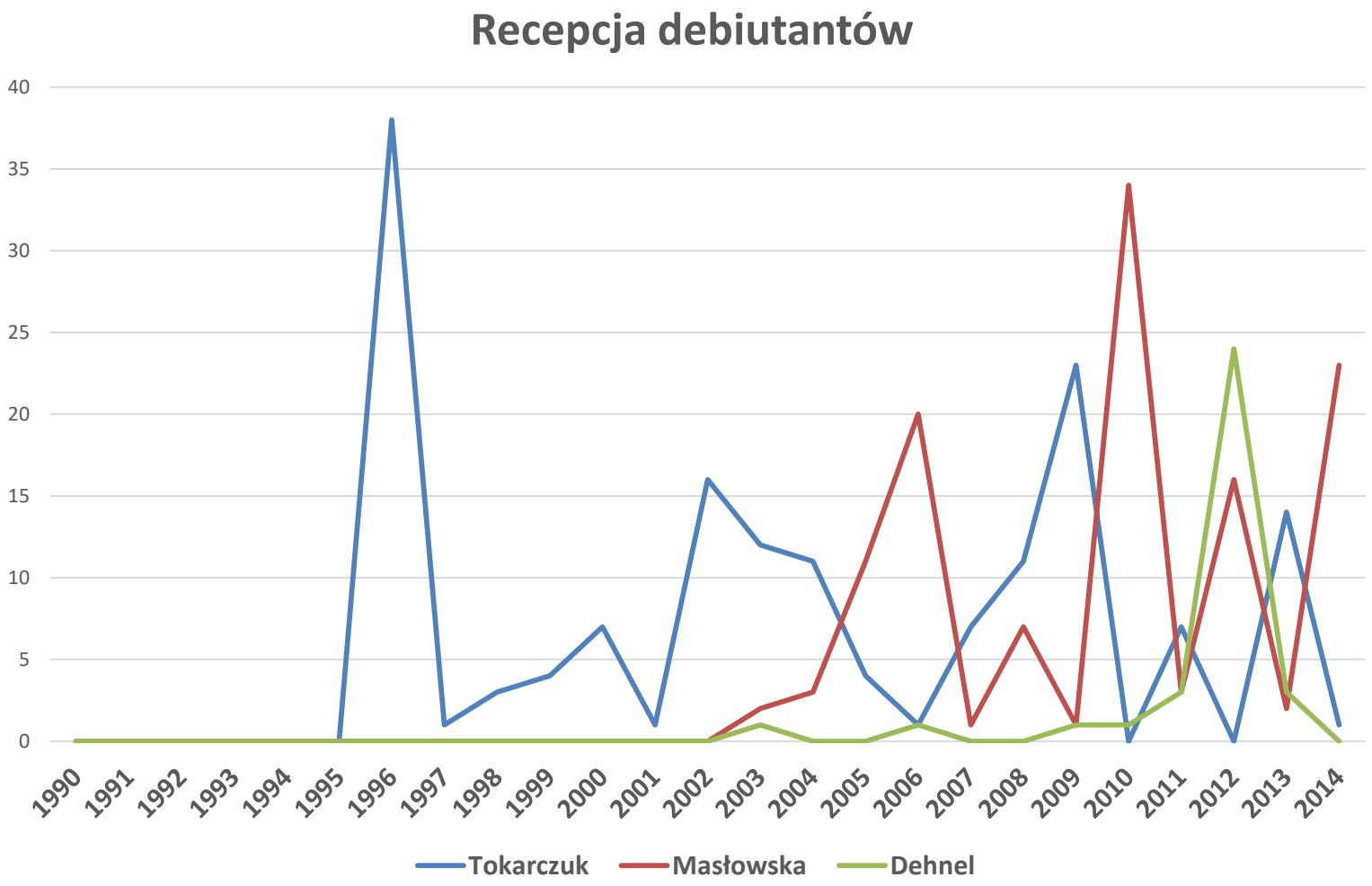
Text mining



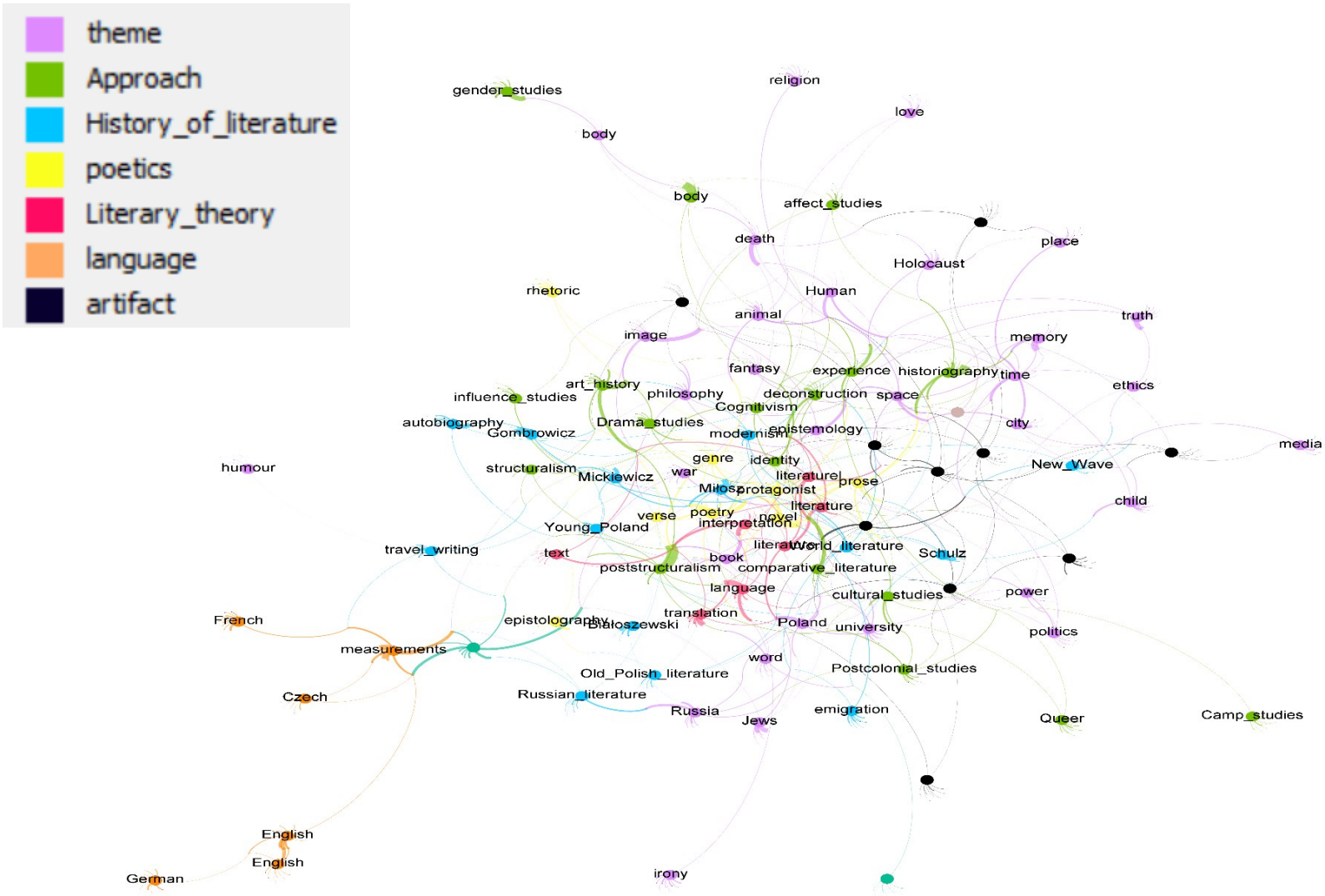
Text mining



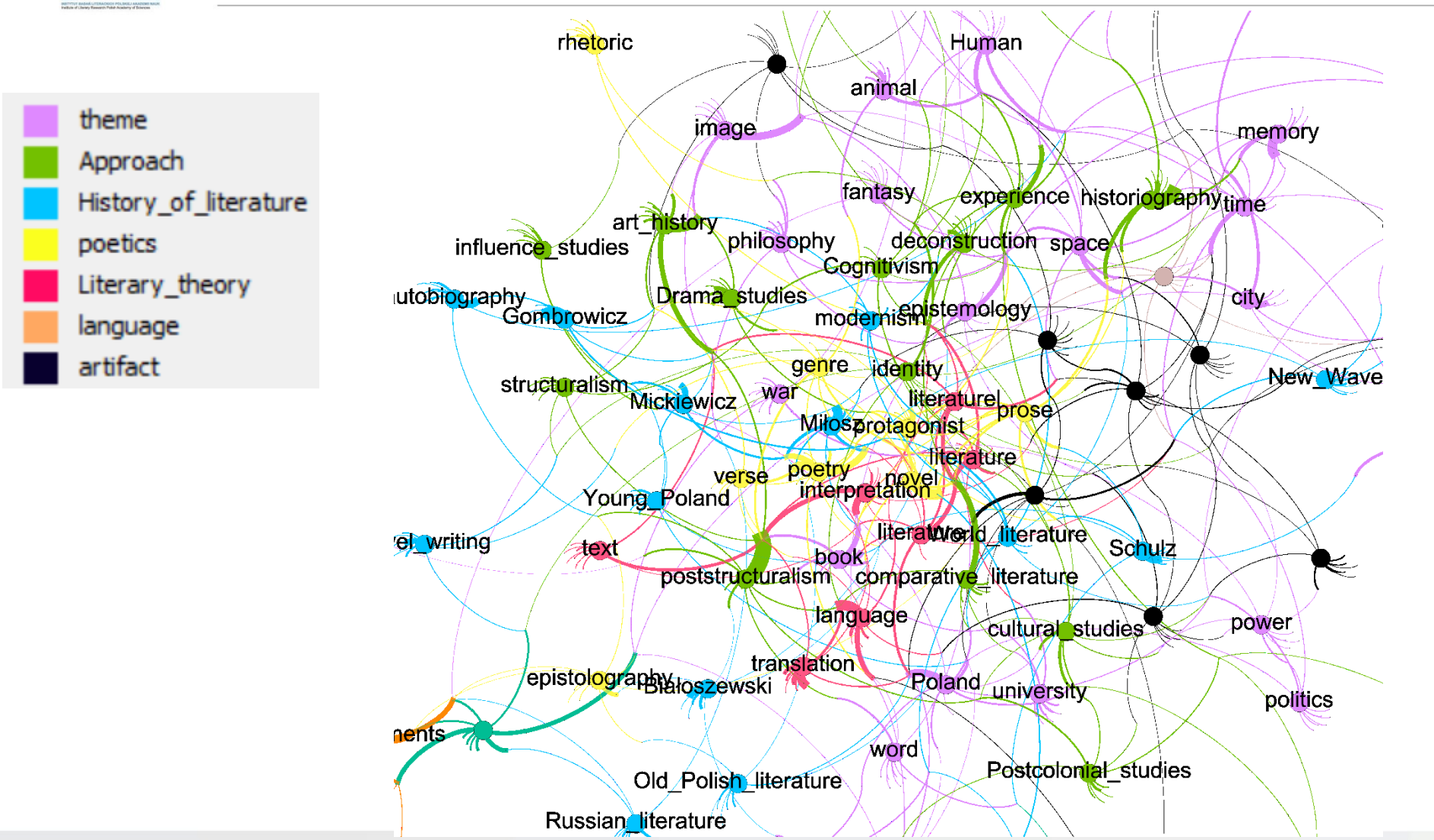
Text mining



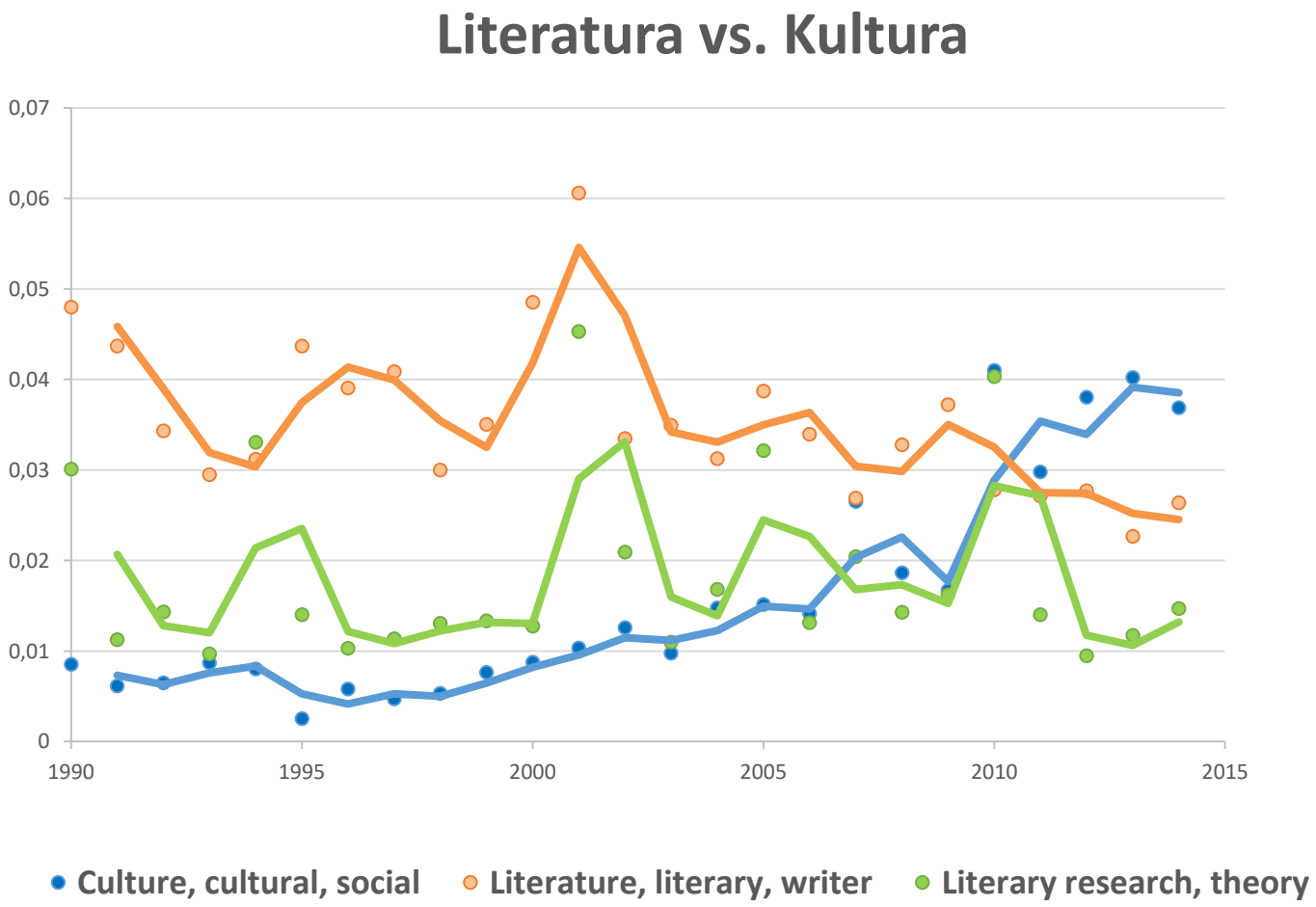
Topic Modelling (Mallet)



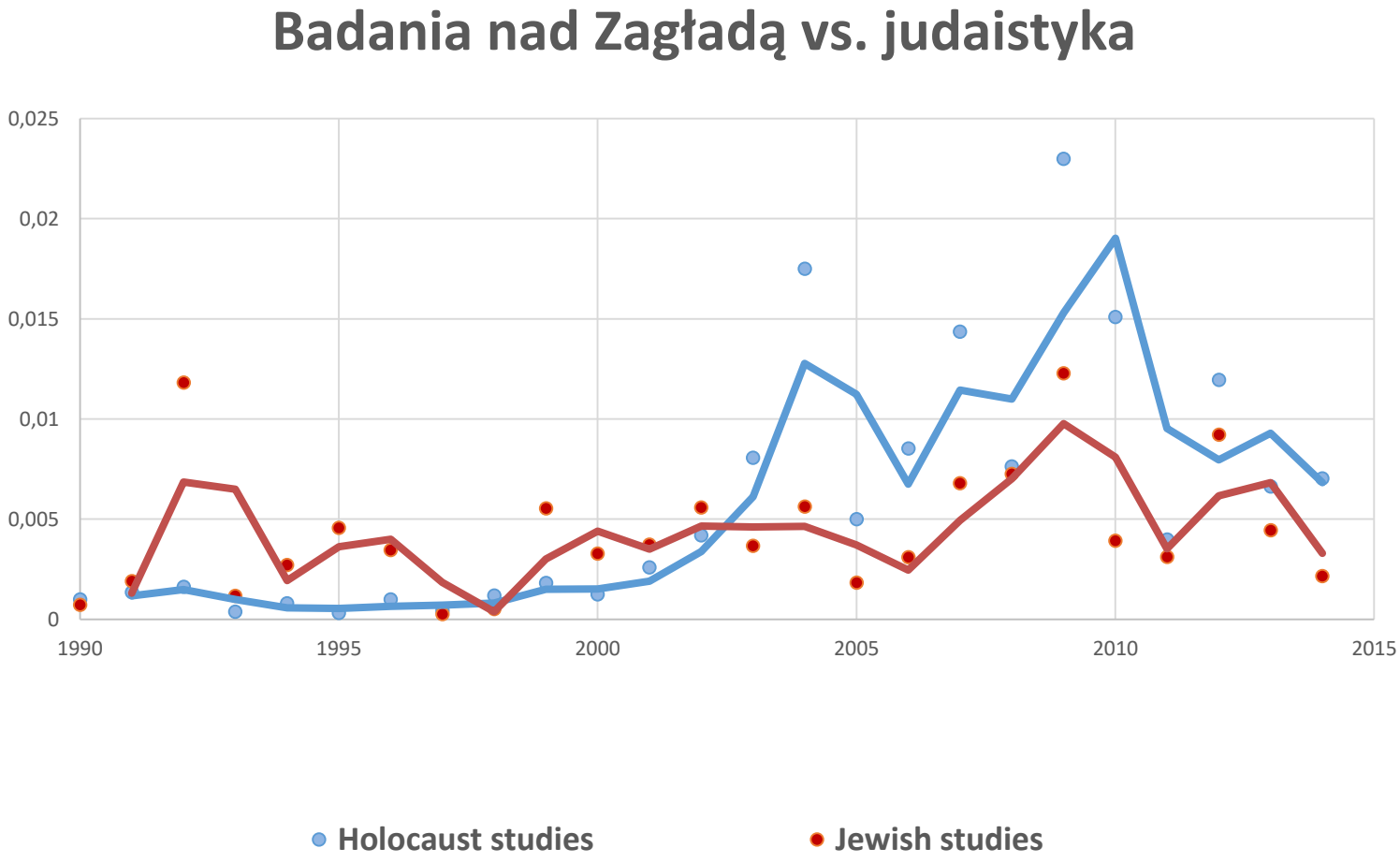
Topic modeling z klasyfikacją



Przemany tematów



Kierunki badawcze



Studium przypadku 4
Dane o literaturze

- Współpraca: Piotr Wciślik
- Maryl, M., Wciślik, P. (2016) “Remediations of Polish Literary Bibliography: Towards a Lossless and Sustainable Retro-Conversion Model for Bibliographical Data.” W *Digital humanities 2016: Conference Abstracts*. Jagiellonian University & Pedagogical University, pp. 621-623. [PDF] [Online]
- Maryl, M. (2015). *Życie literackie w sieci. Pisarze, instytucje i odbiorcy wobec przemian technologicznych*, Warszawa: Wydawnictwo IBL, 469 stron. [PDF]

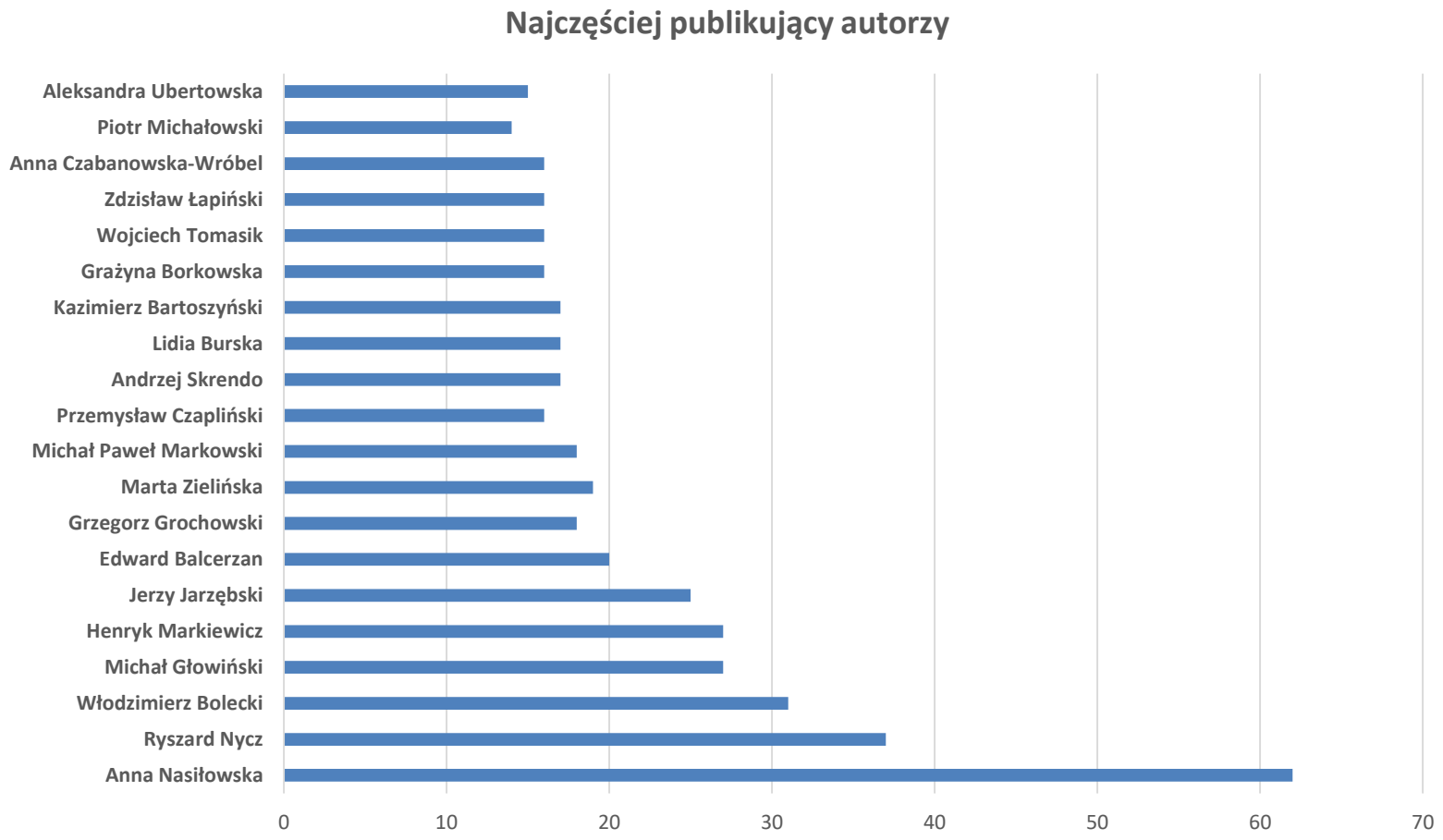
Studium przypadku: Bibliografia *Tekstów Drugich*

Materiał:

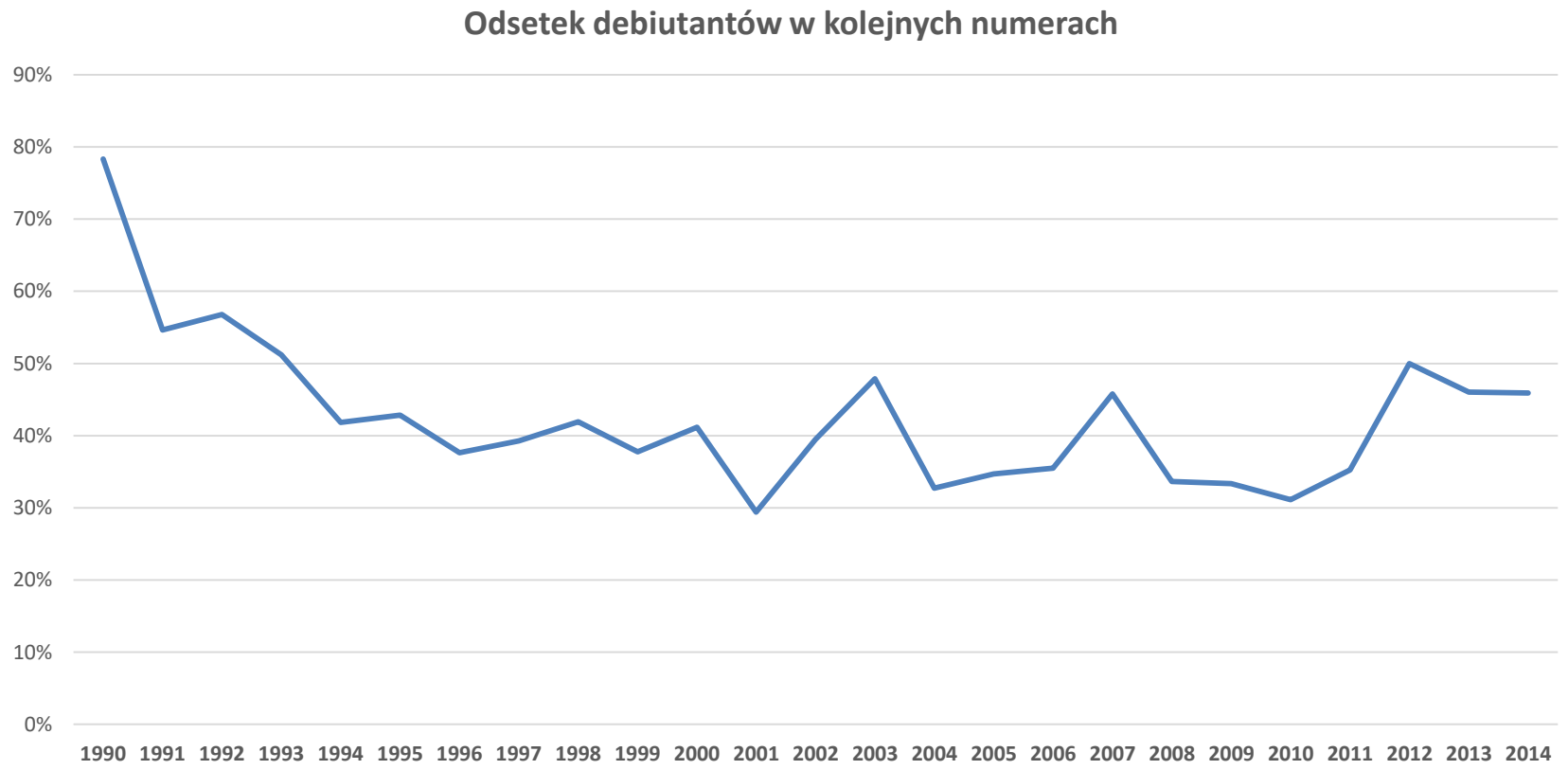
- **Bibliografia ze strony czasopisma (Autor, tytuł, tłumacz, streszczenie) (1990-2015)**
- **Bibliografia z Repozytorium Cyfrowego Instytutów Naukowych (Cytowania) (1990-2012)***

* Dziękuję Marcinowi Helińskiemu (PCSS) za pomoc w pozyskaniu danych

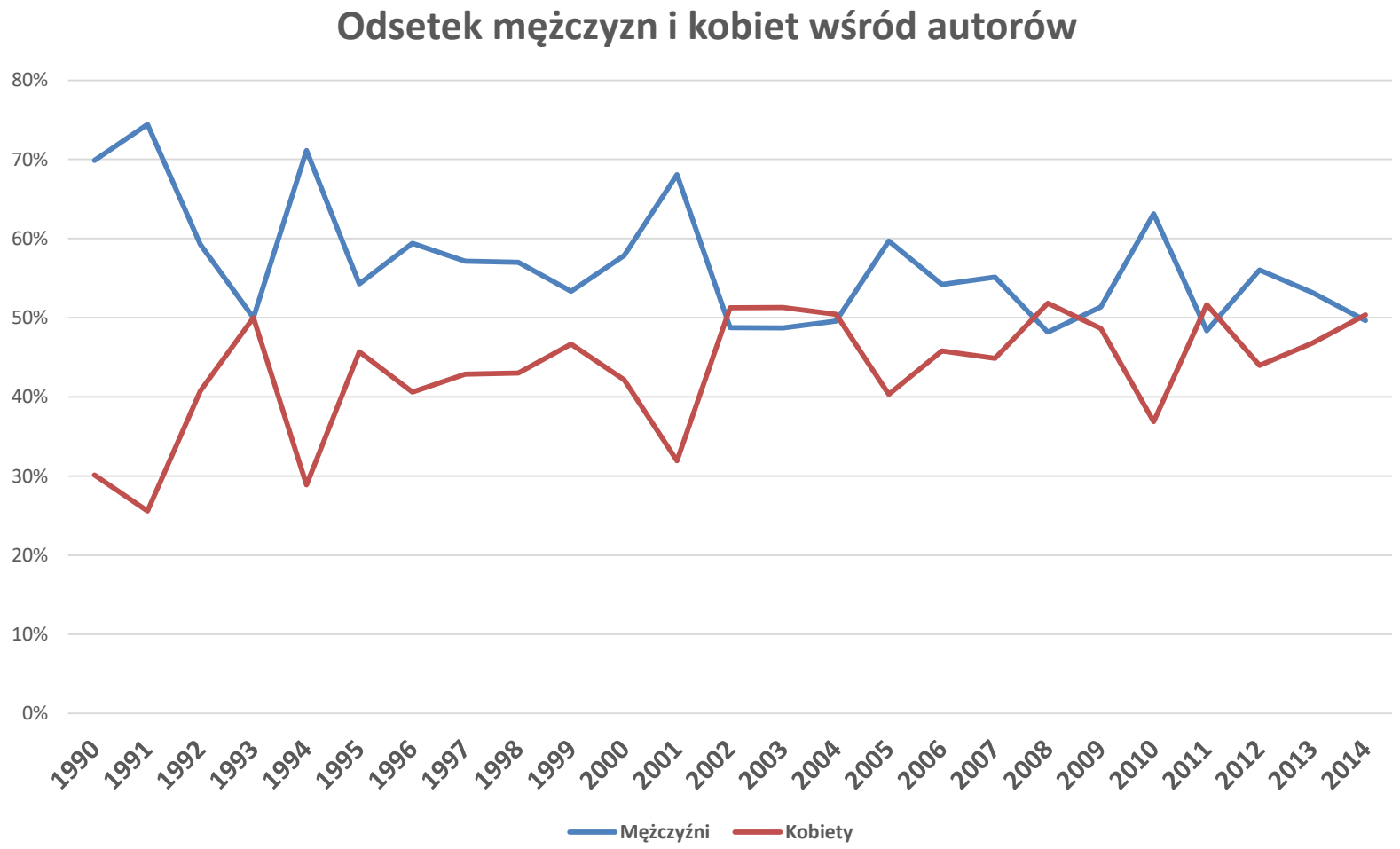
Analiza danych bibliograficznych



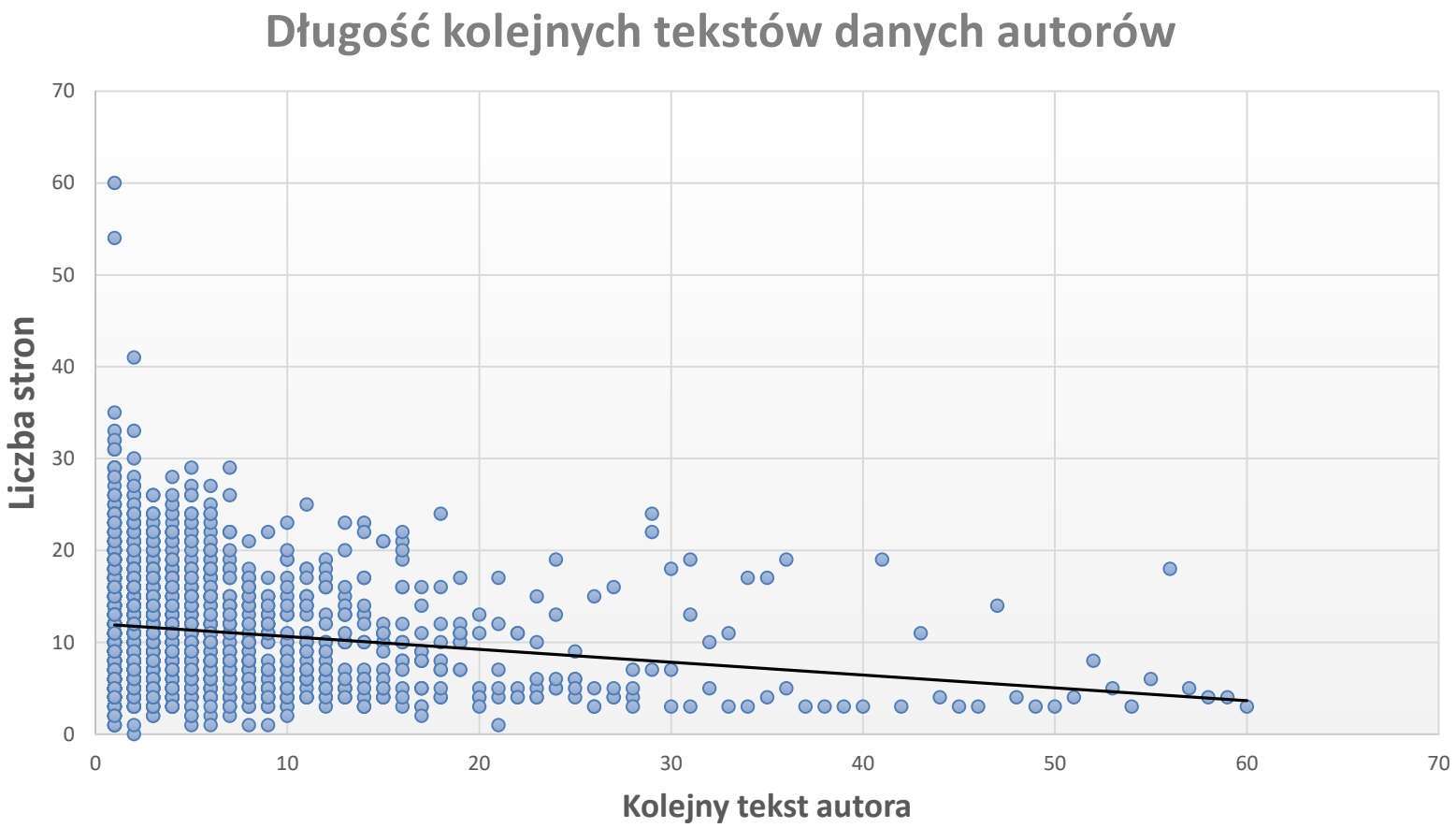
Dane bibliograficzne



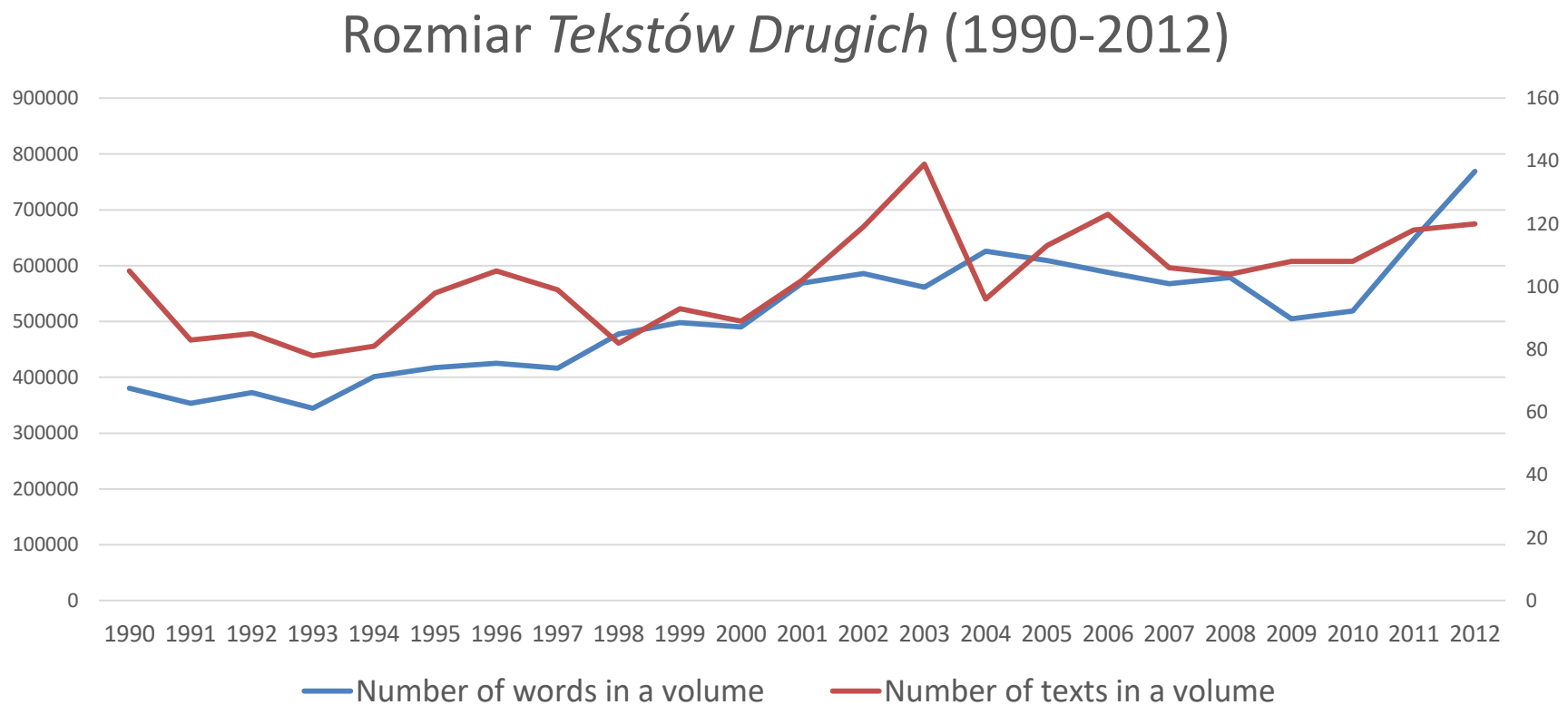
Dane bibliograficzne



Dane bibliograficzne



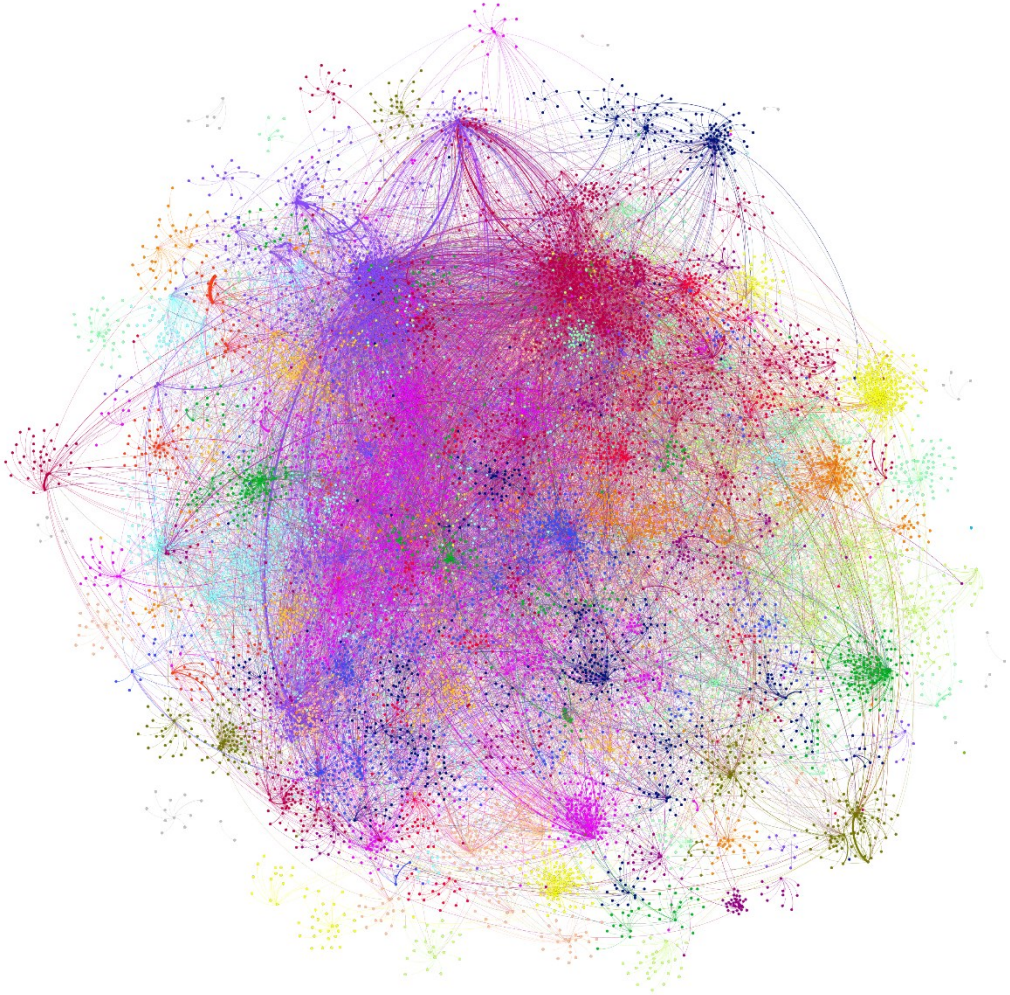
Analiza metadanych



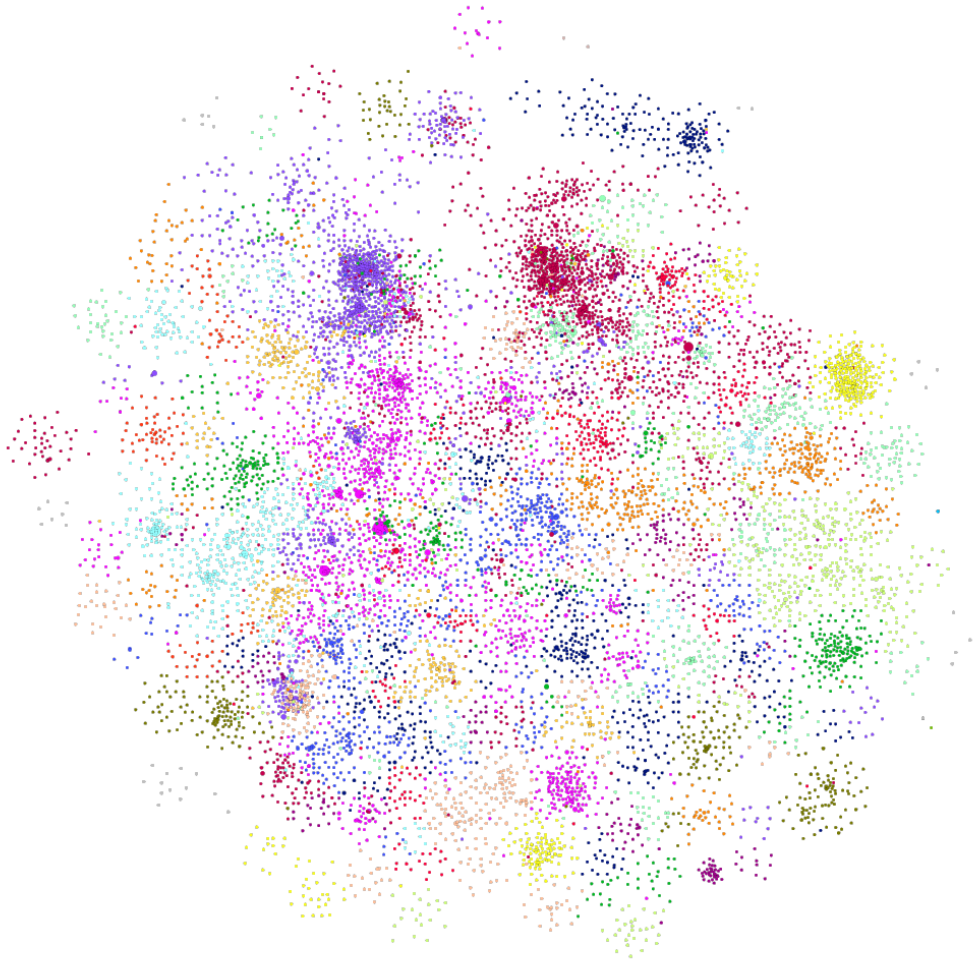
Cytowania

100 najczęściej cytowanych autorów (via Wordle)

Sieć relacji
(dane bibliograficzne)

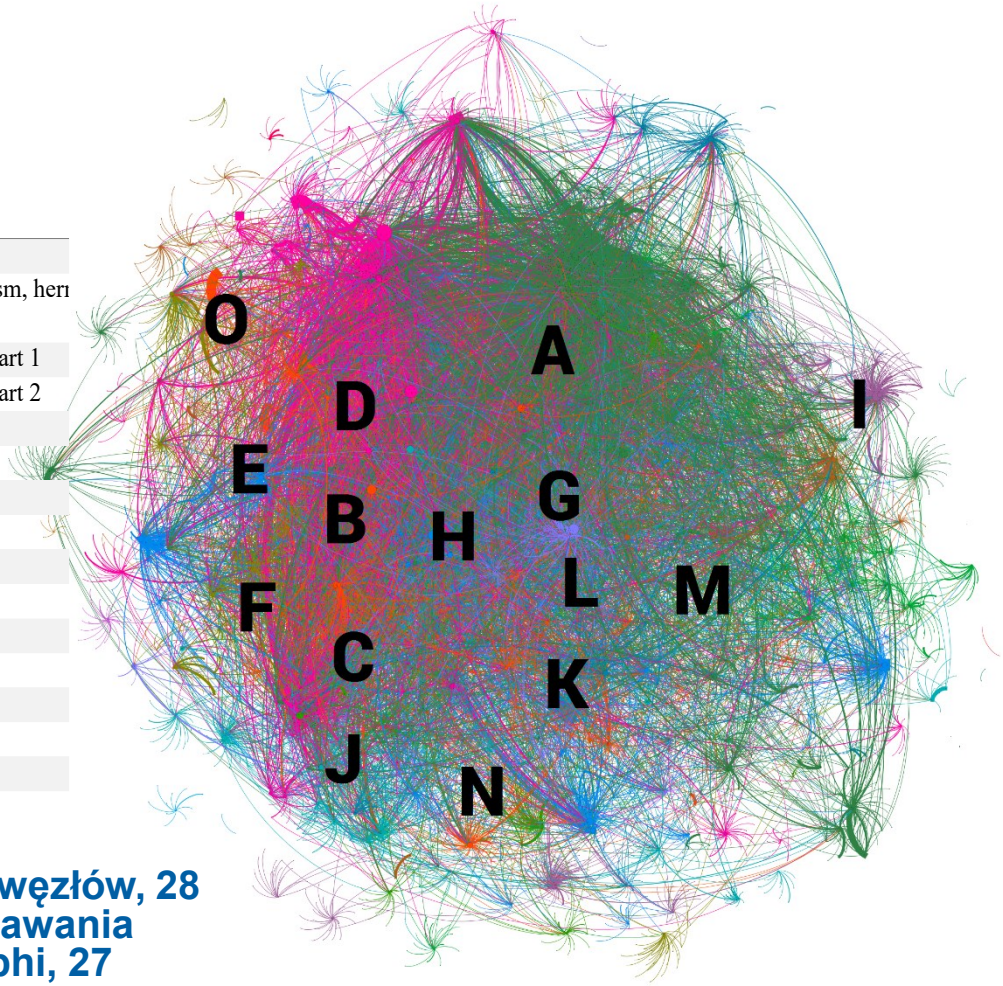


Sieć relacji bez krawędzi (dane bibliograficzne)

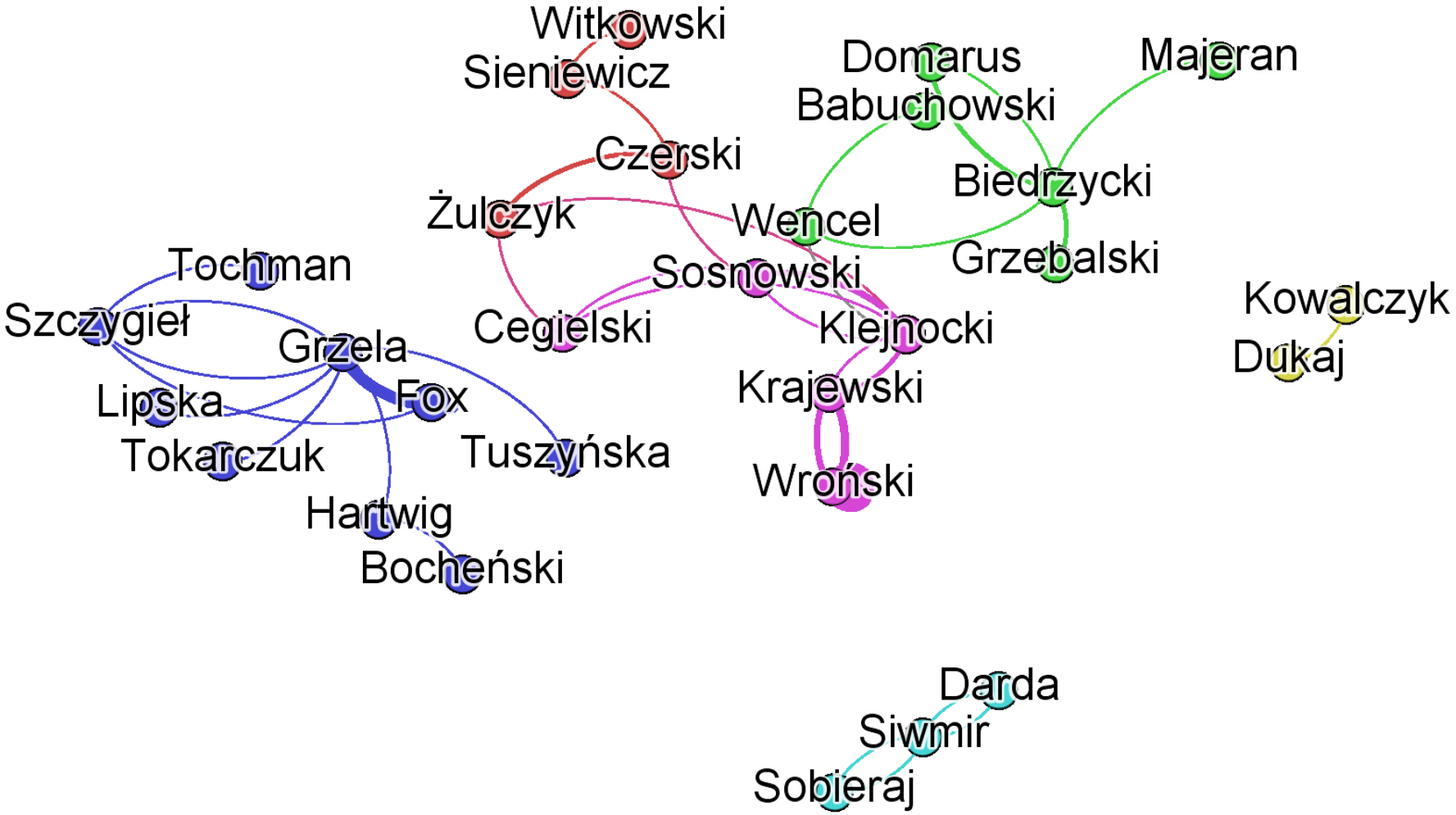


Analiza metadanych

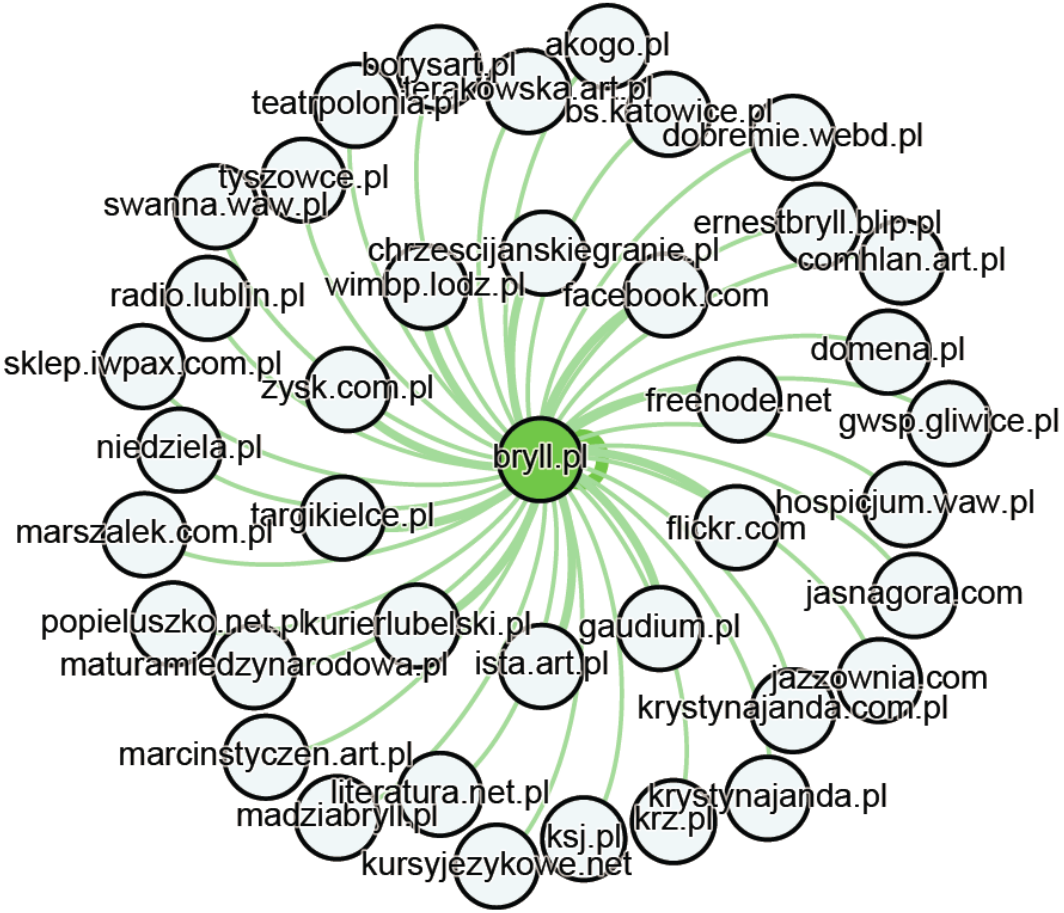
CLUSTER	NUMBER OF AUTHORS	% (N=10543)	TOPIC
A	2510	23.81%	Poststructuralism, philosophy, literaturę
B	1284	12.18%	Theory of literature and culture (structuralism, herı sociology of literature)
C	928	8.80%	History of 20th.-century Polish Literature part 1
D	912	8.65%	History of 20th.-century Polish Literature part 2
E	748	7.09%	Modernism, culture and art.
F	683	6.48%	Romanticism
G	535	5.07%	Holocaust and testimony
H	439	4.16%	Literary criticism, comparative literaturę
I	423	4.01%	Neurosemiotics, darawinism
J	404	3.83%	Feminism
K	387	3.67%	Poetics, versology
L	387	3.67%	Postcolonialism
M	347	3.29%	Old-Polish Literature
N	293	2.78%	Formalism and Prague School
O	220	2.09%	Darvinism
OTHER	43	0.41%	[12 clusters with rare authors]



Sieć współcytowań: ukierunkowana, 10 191 węzłów, 28 613 krawędzi. Zastosowano metodę rozpoznawania skupień Louvain oraz layout OpenOrd w Gephi, 27 clusters


Witkowski
Sieniewicz
Domarus
Majeran
Babuchowski
Czerski
Biedrzycki
Żulczyk
Wencel
Grzebalski
Tochman
Sosnowski
Kowalczyk
Szczygieł
Cegielski
Klejnocki
Dukaj
Grzela
Krajewski
Lipska
Fox
Tokarczuk
Tuszyńska
Wroński
Hartwig
Bocheński
Darda
Siwmir
Sobieraj

Wizualizacja sieci pisarza (analiza linków)



W stronę literackich *Big Data*

- Korzystanie z danych bibliograficznych, adnotacji i statystyk tekstowych (text mining)
- Metody ilościowe wspomagające interpretację jakościową (badania zintegrowane)
- Lepsze dane → lepsze odpowiedzi
- Więcej danych → więcej problemów
- Praca na danych, które powstały dla celów informacji naukowej, nie badań (np. *casus* adnotacji)
- Cel: stworzenie bazy danych zdolnej udzielać odpowiedzi na uprzednio niezaplanowane pytania

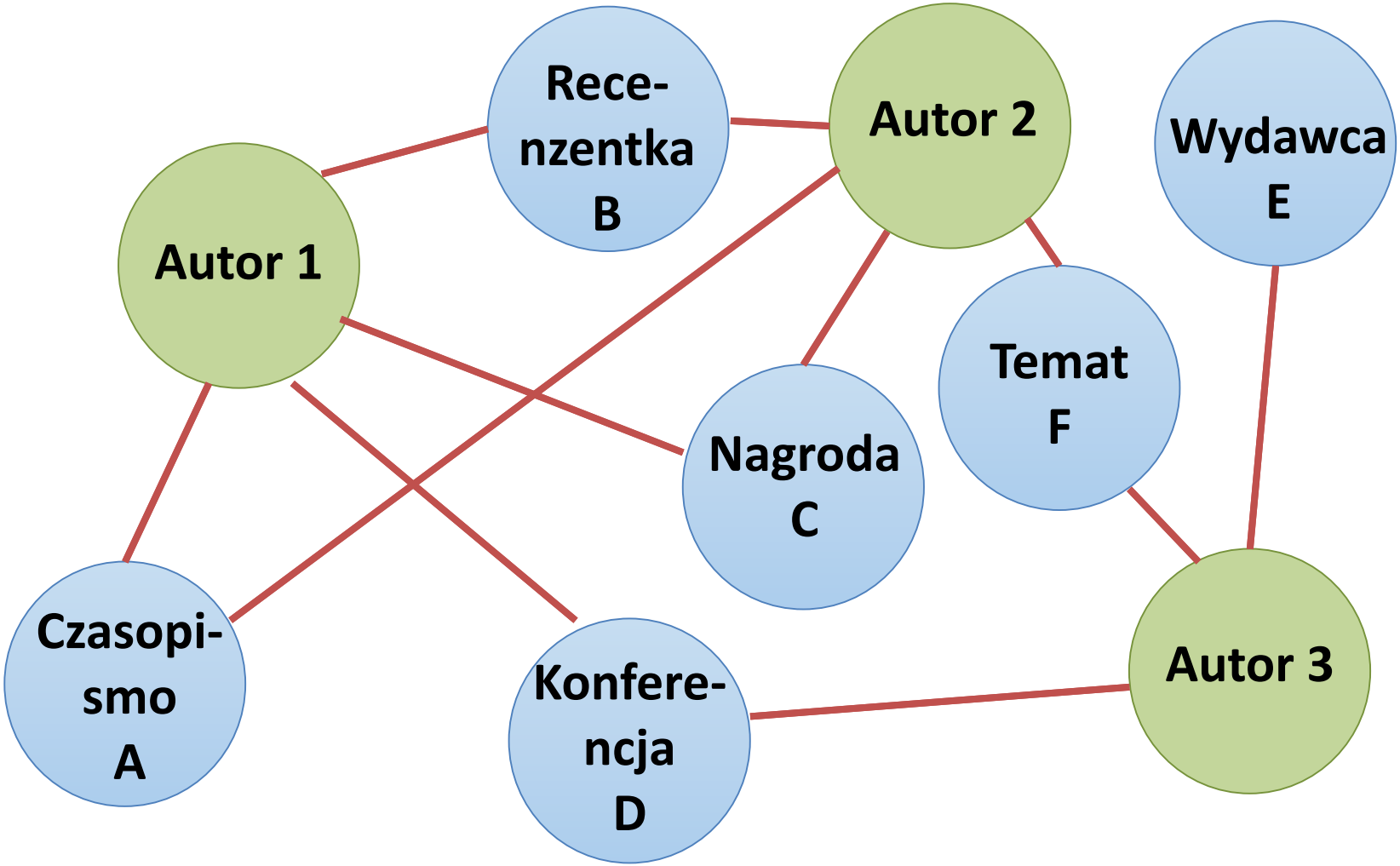
PBL_lab

- **Projekt:**
  - **Polska Bibliografia Literacka – laboratorium wiedzy o współczesnej kulturze polskiej;**
  - **3-letni projekt realizowany ze środków Narodowego Programu Rozwoju Humanistyki (2015 – 2018) (okrojony o 2/3 w stosunku do planowanego zakresu).**
  - **Partner informatyczny: Poznańskie Centrum Superkomputerowo-Sieciowe przy IChB PAN**

Cele:

- **Stworzenie bazy danych potencjalnie obejmującej lata 1939-2002 (ok. 3 miliony rekordów)**
- **Scalenie bazy online z rocznikami archiwalnymi i bibliografiami pochodnymi (II WŚ; Drugi Obieg)**
- **Połączenie zapisów z elementami chmury *Linked Open Data.***
- **Stworzenie prototypów narzędzi do prowadzenia badań**

PBL lab
Narzędzie badawcze



Morał

- Teksty i dane jako narzędzia pozyskiwania nowej wiedzy
- Potrzebujemy więcej danych!
- Narzędzia ułatwiają, a nie zastępują interpretację
- „Trzecia fala humanistyki cyfrowej” – krytyczny namysł nad epistemologią narzędzi

Dziękuję za uwagę

Maciej.Maryl@ibl.waw.pl

ORCID ID 0000-0002-2639-041X

@maciejmaryl

maryl.org

chc.ibl.waw.pl