# Otwarte dane bibliograficzne (open bibliographic data) i ich wykorzystanie w krajowych i międzynarodowych projektach z zakresu humanistyki cyfrowej

**Dorota Siwecka**
Instytut Nauk o Informacji i Mediach
Uniwersytet Wrocławski
dorota.Siwecka@uwr.edu.pl

VI POMORSKA KONFERENCJA OPEN SCIENCE – POLITYKI I INFRASTRUKTURY DLA OTWARTEJ NAUKI

# PLAN WYSTĄPIENIA

Krótko o humanistyce cyfrowej

Odrobinę historii

Badania oparte na DB

Bibliographic Data Science

Dane bibliograficzne danymi badawczymi

Infrastruktura

# HUMANISTYKA CYFROWA

„zbiór praktyk poznawczych i zasobów informacji, będący efektem przeniesienia do sfery cyfrowej i wzbogacenia o nowe funkcjonalności praktyk poznawczych humanistyki opartej na druku” (A. Pawłowski)

Badanie wielkich zbiorów danych (tekstowych, dźwiękowych, wizualnych)

Narzędzia cyfrowe ułatwiają analizę tych danych oraz wnioskowanie na ich podstawie

- NLP
- Text minig
- Data mining
- Narzędzia do wizualizacji danych
- Rzutowanie danych geograficznych na mapy
- ….

# DANE BIBLIOGRAFICZNE

## Jakie elementy można badać?

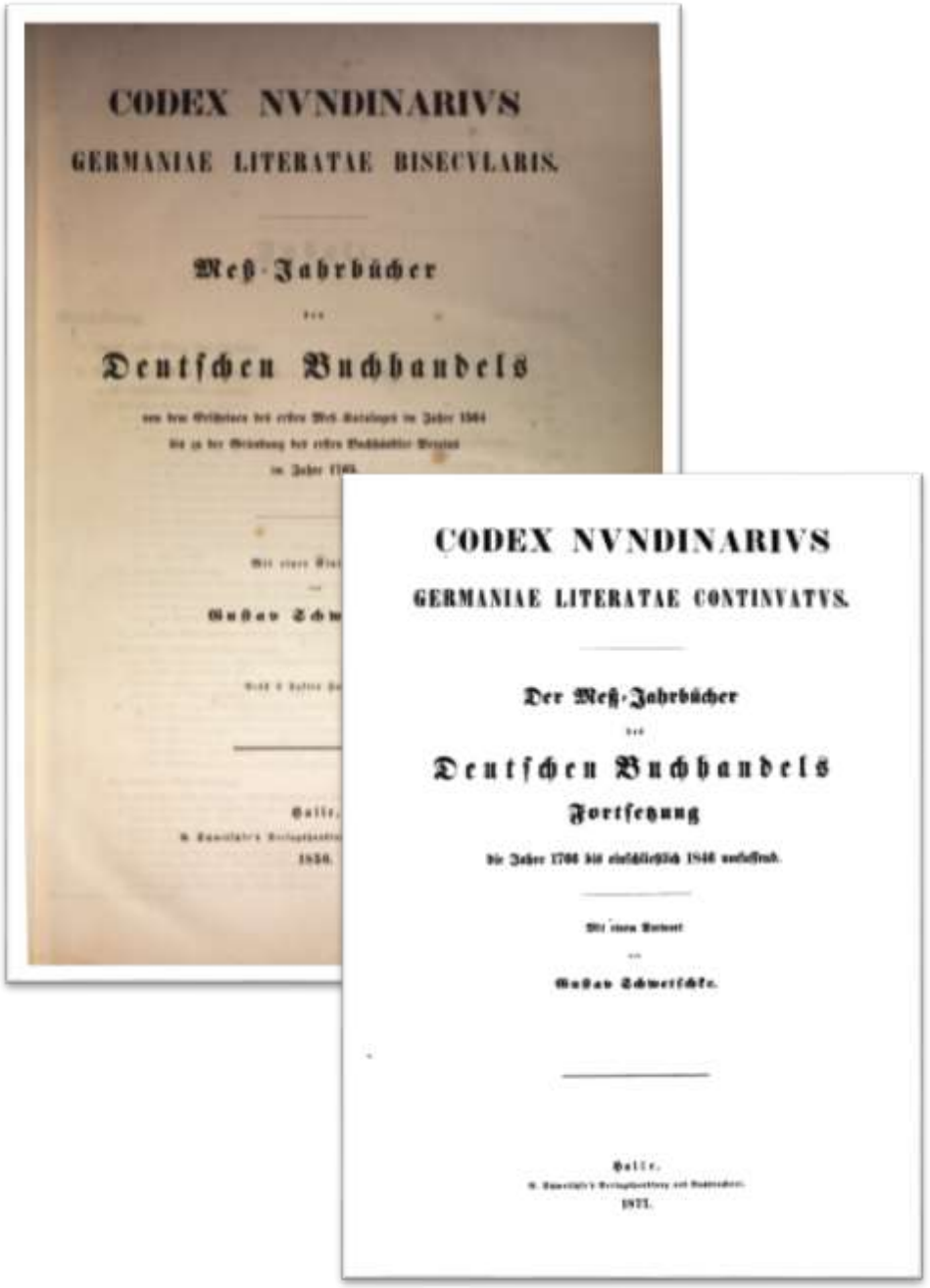
- Metadane – w zależności od formatu i schematu metadanych:

| Elementy związane z formalnymi cechami dokumentów | Elementy związane z rzeczowymi cechami dokumentów (tematem) |
|---|---|
| •Autor, tytuł, miejsce wydania, rok wydania, wydawca | •Hasła przedmiotowe, deskryptory, słowa kluczowe |

# ODROBINA HISTORII

Gustav Schwetschke. (**1850–1877**). *Codex nundinarius Germaniae literatae continuatus…* 1546–1846

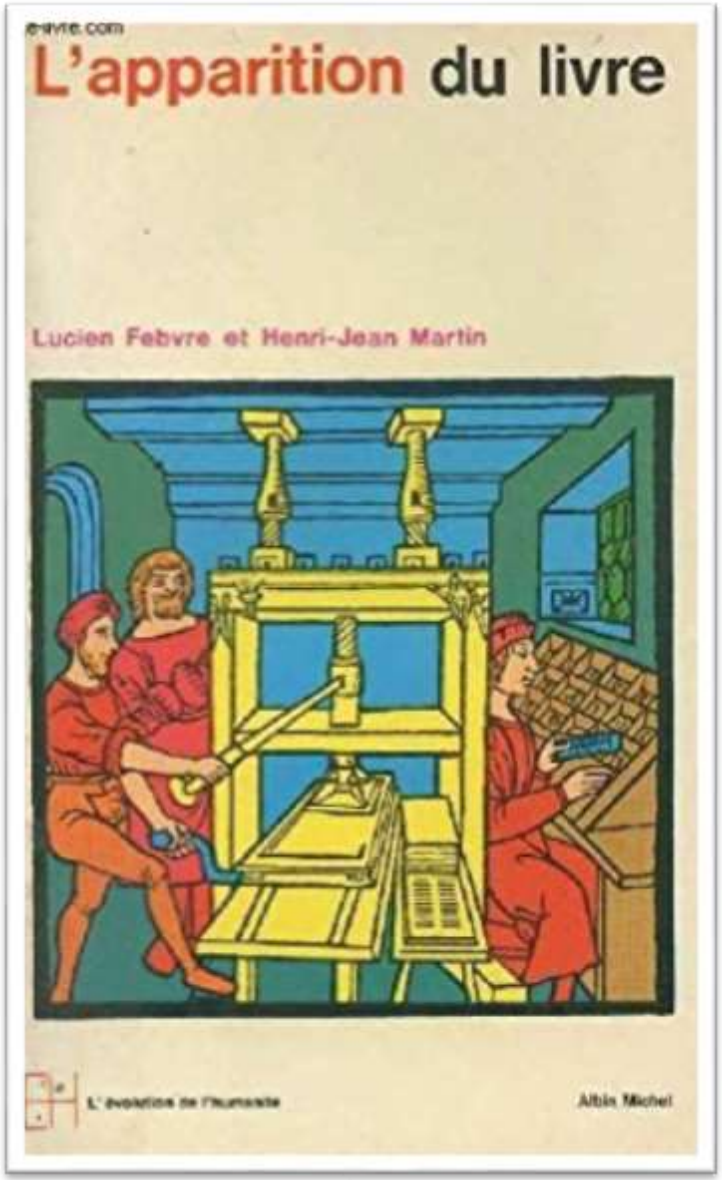
- geograficzny i liczbowy rozkład drukarstwa niemieckiego XVI–XIX w.
- podstawa: katalogi targów książki w Lipsku i Frankfurcie nad Menem



# ODROBINA HISTORII

Febvre, L., & Martin, H. J. (**1958**). *L'Apparition du Livre*

- książka jako nośnik idei i towar
- podstawa: katalogi inkunabułów, retrospektywne bibliografie narodowe i specjalne
- Polskie tłumaczenie: *Narodziny książki*



# ODROBINA HISTORII

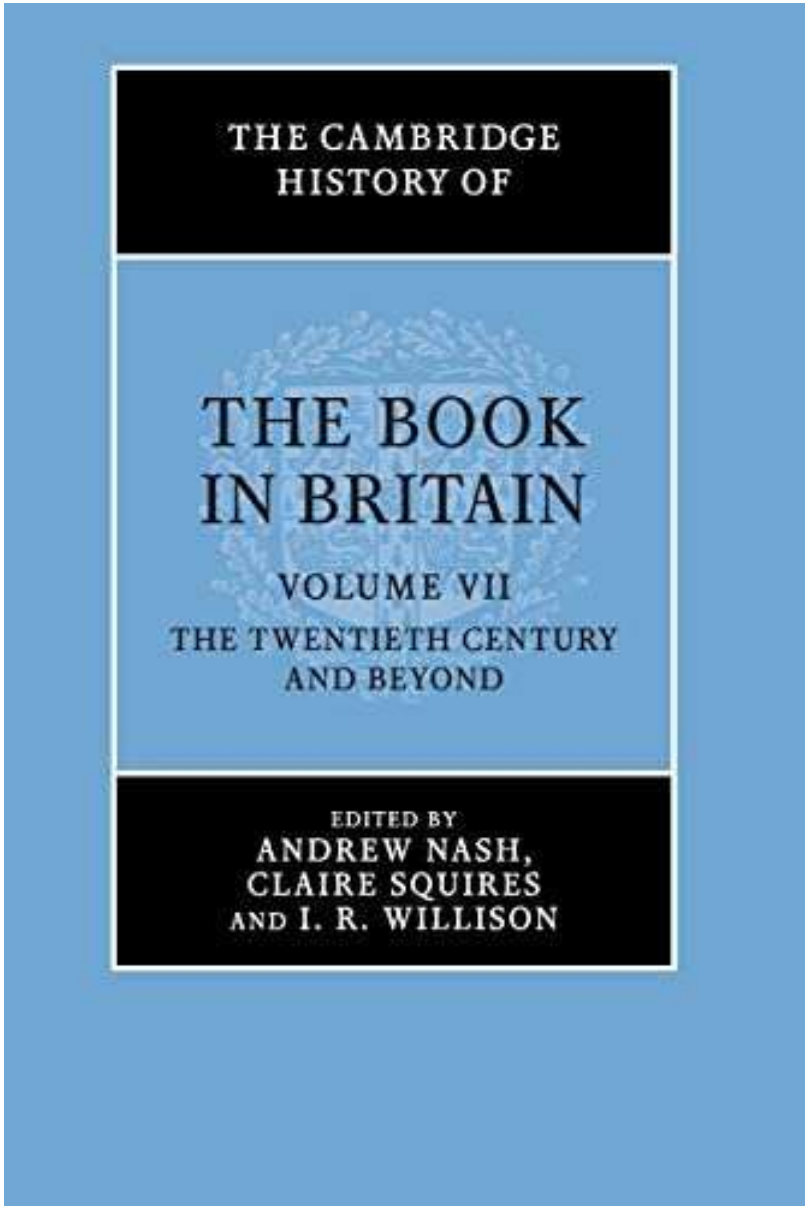
Czarnowska, M. (**1967**). *Ilościowy rozwój polskiego ruchu wydawniczego 1501-1965*

- podstawa: *Bibliografia polska* Estreichera, polska bieżąca bibliografia narodowa („PB”, „UWD”)

# WSPÓŁCZESNE OPRACOWANIA

Gameson, R. et al. (Red.). (1999). *The Cambridge history of the book in Britain* (T. 1– 7).

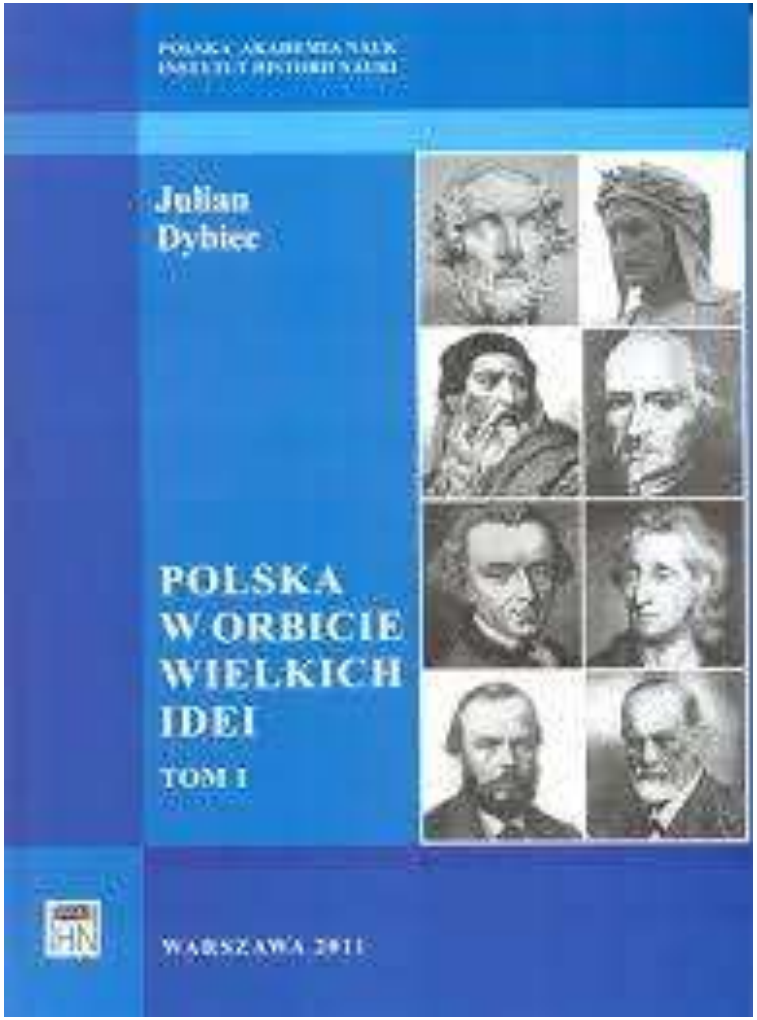
- dzieje książki na wyspach brytyjskich od 400 r. do końca XX w.
- podstawa: różnorodne materiały bibliograficzne



# WSPÓŁCZESNE OPRACOWANIA

Dybiec, J. (2011). *Polska w orbicie wielkich idei: Polskie przekłady obcojęzycznego piśmiennictwa 1795-1918*

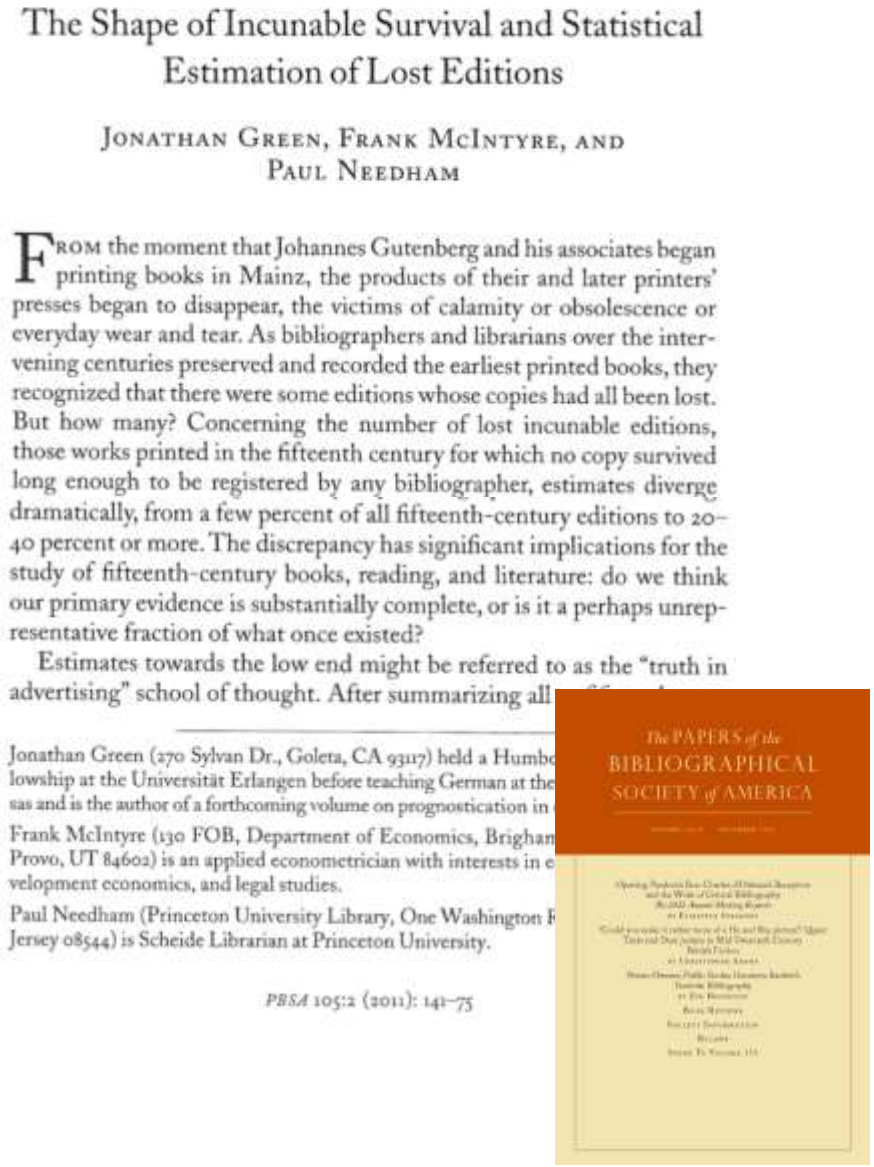
- tłumaczenia obcojęzycznych dzieł naukowych
- podstawa: *Bibliografia polska* Estreichera, polska bieżąca bibliografia narodowa



# BADANIA OPARTE NA DANYCH BIBLIOGRAFICZNYCH

Green, J., McIntyre, F., & Needham, P. (2011). The Shape of Incunable Survival and Statistical Estimation of Lost Editions. https://doi.org/10.1086/680773

- Analiza stanu zachowania dawnej produkcji drukarskiej
- podstawa: zdigitalizowane bibliografie, bazy inkunabułów i starych druków



The Shape of Incunable Survival and Statistical Estimation of Lost Editions

Jonathan Green, Frank McIntyre, and Paul Needham

From the moment that Johannes Gutenberg and his associates began printing books in Mainz, the products of their and later printers' presses began to disappear, the victims of calamity or obsolescence or everyday wear and tear. As bibliographers and librarians over the intervening centuries preserved and recorded the earliest printed books, they recognized that there were some editions whose copies had all been lost. But how many? Concerning the number of lost incunable editions, those works printed in the fifteenth century for which no copy survived long enough to be registered by any bibliographer, estimates diverge dramatically, from a few percent of all fifteenth-century editions to 20–40 percent or more. The discrepancy has significant implications for the study of fifteenth-century books, reading, and literature: do we think our primary evidence is substantially complete, or is it a perhaps unrepresentative fraction of what once existed?

Estimates towards the low end might be referred to as the "truth in advertising" school of thought. After summarizing all

Jonathan Green (270 Sylvan Dr., Goleta, CA 93117) held a Humbo lowship at the Universität Erlangen before teaching German at the sas and is the author of a forthcoming volume on prognostication in

Frank McIntyre (130 FOB, Department of Economics, Brighan Provo, UT 84602) is an applied econometrician with interests in e velopment economics, and legal studies.

Paul Needham (Princeton University Library, One Washington Jersey 08544) is Scheide Librarian at Princeton University.

*PBSA* 105:2 (2011): 141–75

# BADANIA O CHARAKTERZE HISTORYCZNYM

Leo Lahti, Jani Marjanen, Hege Roivainen, Mikko Tolonen (2019)

- **ewolucja formatów książek** wydawanych w Europie między XVI a XIX w. oraz **proces upiśmiennienia** w Europie, ze szczególnym uwzględnieniem języka szwedzkiego i fińskiego
- ponad 6 milionów rekordów bibliograficznych
- źródła danych:
  - szwedzka bibliografia narodowa
  - fińska bibliografia narodowa
  - ESTC (English Short-Title Catalogue)
  - HPBD (Heritage of the Printed Book Database – Baza Dziedzictwa Książki Drukowanej)

CATALOGING & CLASSIFICATION QUARTERLY

Routledge

OPEN ACCESS

**Bibliographic Data Science and the History of the Book (c. 1500–1800)**

Leo Lahti[a], Jani Marjanen[b], Hege Roivainen[b] and Mikko Tolonen[b]

[a]Department of Mathematics and Statistics, University of Turku, Finland; [b]Helsinki Computational History Group, Department of Digital Humanities, University of Helsinki, Finland

ABSTRACT

National bibliographies have been identified as a crucial resource for historical research on the publishing landscape, but using them requires addressing challenges of data quality, completeness, and interpretation. We call this approach bibliographic data science. In this article, we briefly assess the development of book formats and the vernacularization process in early modern Europe. The work undertaken paves the way for more extensive integration of library catalogs to map the history of the book.

ARTICLE HISTORY

KEYWORDS

**Introduction**

Library catalogs are essential tools in information science, and their utilization has been greatly advanced by digitalization.[1] The need to manage and organize the ever-increasing body of digital information has motivated the development of new concepts and technologies, such as Linked Data, which was first introduced some 20 years ago and has been on the agenda of most National Libraries since then. During the last decade, the concept of Linked Open Data (LOD) emerged to emphasize the importance of open licensing of the data resources. Metadata collections of published material that different libraries hold are particularly suitable for interlinking and enriching with different semantic layers.[2] LOD represents a crucial step in taking full advantage of digital resources through the integration of web sources and open, reusable metadata and its enrichment.[3]

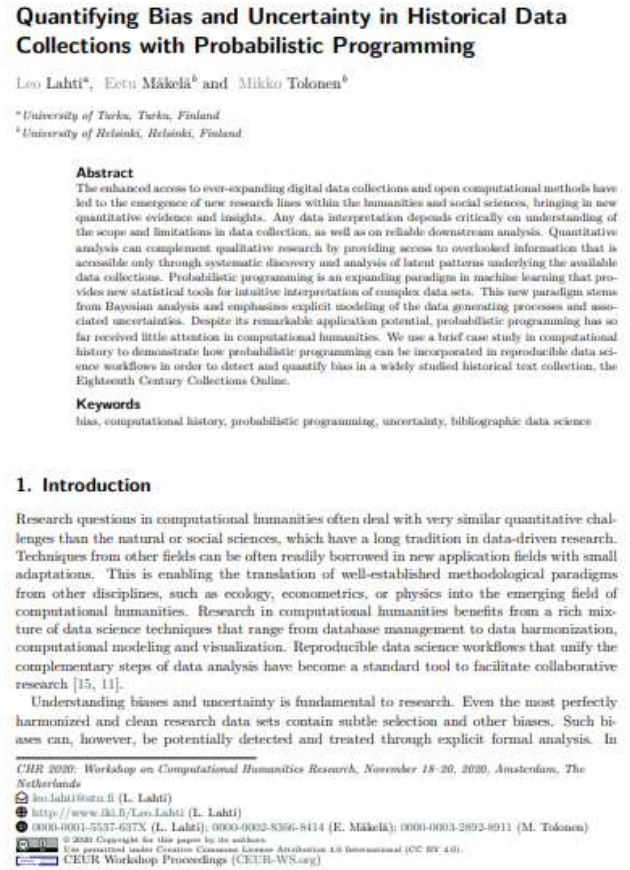
National bibliographies have been traditionally used as a tool for information retrieval. This article demonstrates our quantitative approach to book history, where bibliographic collections are considered as research material, rather than a mere retrieval tool. A key feature in this work is

CONTACT Mikko Tolonen

# BADANIA O CHARAKTERZE HISTORYCZNYM

Leo Lahti, Eetu Mäkelä, Mikko Tolonen (2020), *Quantifying Bias and Uncertainty in Historical Data Collections with Probabilistic Programming*

- wykrycie poziomu dostępności pełnych tekstów w ECCO zarejestrowanych w bibliografii ESTC
- 227 tys. rekordów
- źródła danych:
  - ESTC (English Short-Title Catalogue) – źródło metadanych
  - ECCO (Eighteenth Century Collection Online) – informacje o dostępności online
  - VIAF – daty życia i płeć autora



**Quantifying Bias and Uncertainty in Historical Data Collections with Probabilistic Programming**

Leo Lahti[a], Eetu Mäkelä[b] and Mikko Tolonen[b]

[a]University of Turku, Turku, Finland
[b]University of Helsinki, Helsinki, Finland

**Abstract**

The enhanced access to ever-expanding digital data collections and open computational methods have led to the emergence of new research lines within the humanities and social sciences, bringing in new quantitative evidence and insights. Any data interpretation depends critically on understanding of the scope and limitations in data collection, as well as on reliable downstream analysis. Quantitative analysis can complement qualitativ research by providing access to overlooked information that is accessible only through systematic discovery and analysis of latent patterns underlying the available data collections. Probabilistic programming is an expanding paradigm in machine learning that provides new statistical tools for intuitive interpretation of complex data sets. This new paradigm stems from Bayesian analysis and emphasizes explicit modeling of the data generating processes and associated uncertainties. Despite its remarkable application potential, probabilistic programming has so far received little attention in computational humanities. We use a brief case study in computational history to demonstrate how probabilistic programming can be incorporated in reproducible data science workflows in order to detect and quantify bias in a widely studied historical text collection, the Eighteenth Century Collections Online.

**Keywords**

bias, computational history, probabilistic programming, uncertainty, bibliographic data science

**1. Introduction**

Research questions in computational humanities often deal with very similar quantitative challenges than the natural or social sciences, which have a long tradition in data-driven research. Techniques from other fields can be often readily borrowed in new application fields with small adaptations. This is enabling the translation of well-established methodological paradigms from other disciplines, such as ecology, econometrics, or physics into the emerging field of computational humanities. Research in computational humanities benefits from a rich mixture of data science techniques that range from database management to data harmonization, computational modeling and visualization. Reproducible data science workflows that unify the complementary steps of data analysis have become a standard tool to facilitate collaborative research [15, 11].

Understanding biases and uncertainty is fundamental to research. Even the most perfectly harmonized and clean research data sets contain subtle selection and other biases. Such biases can, however, be potentially detected and treated through explicit formal analysis. In

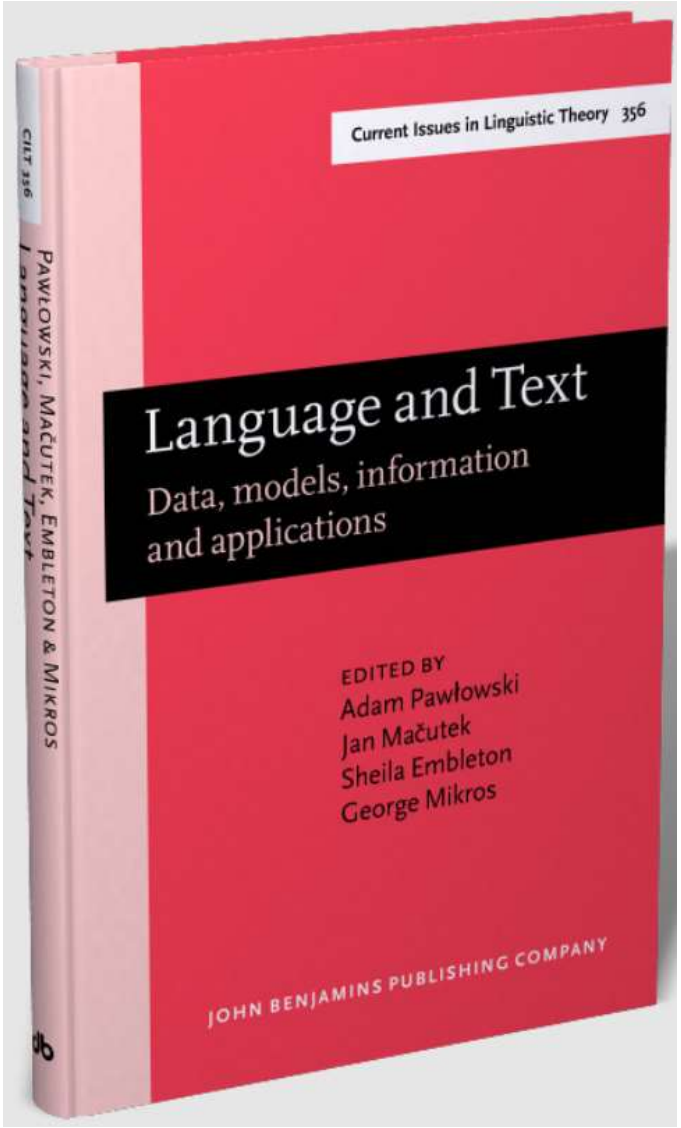
*CHR 2020: Workshop on Computational Humanities Research, November 18–20, 2020, Amsterdam, The Netherlands*
leo.lahti@utu.fi (L. Lahti)
http://www.iki.fi/Leo.Lahti (L. Lahti)
0000-0001-5537-637X (L. Lahti); 0000-0002-8366-8414 (E. Mäkelä); 0000-0003-2892-8911 (M. Tolonen)
© 2020 Copyright for this paper by its authors.
Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).
CEUR Workshop Proceedings (CEUR-WS.org)

280

# BADANIA O CHARAKTERZE LINGWISTYCZNYM

Adam Pawłowski, Krzysztof Topolski, Elżbieta Herden (2021)

- Badania nad tytułami publikacji i porównanie z Narodowym Korpusem Języka Polskiego → obraz konstrukcji językowych w tytułach współczesnych publikacji książkowych
- Źródło: Polska bieżąca bibliografia narodowa „Przewodnik Bibliograficzny”
- 553 000 rekordów z lat 1997-2017 w formacie MARC21
- Metody lingwistyczne wspomagane eksploracją tekstu (*text mining*) – analiza stylostatystyczna, metody NLP, wykorzystanie WCRF Tagger, WCRFT2 do lematyzacji tekstów tytułów oraz oznaczania części mowy



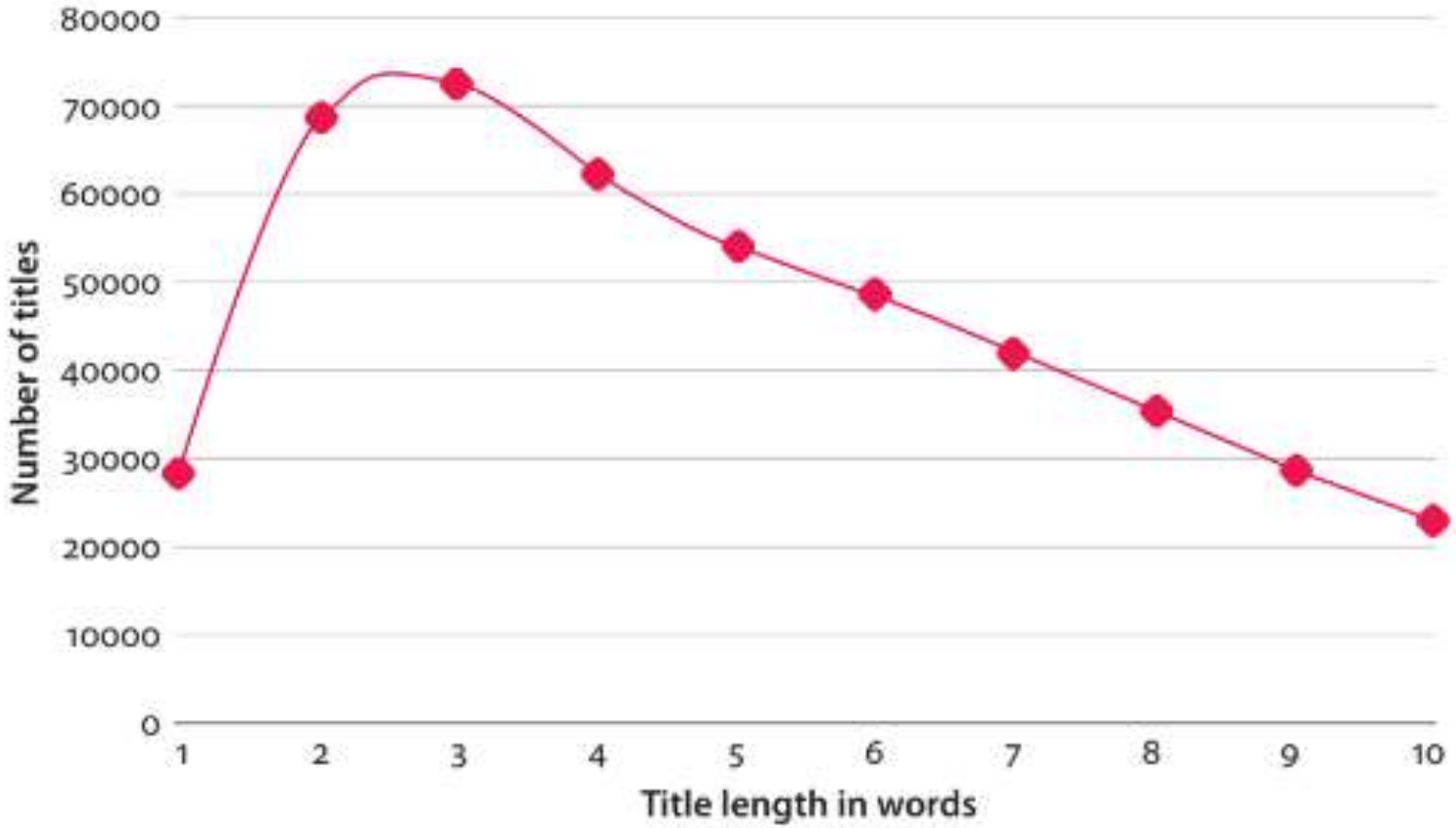


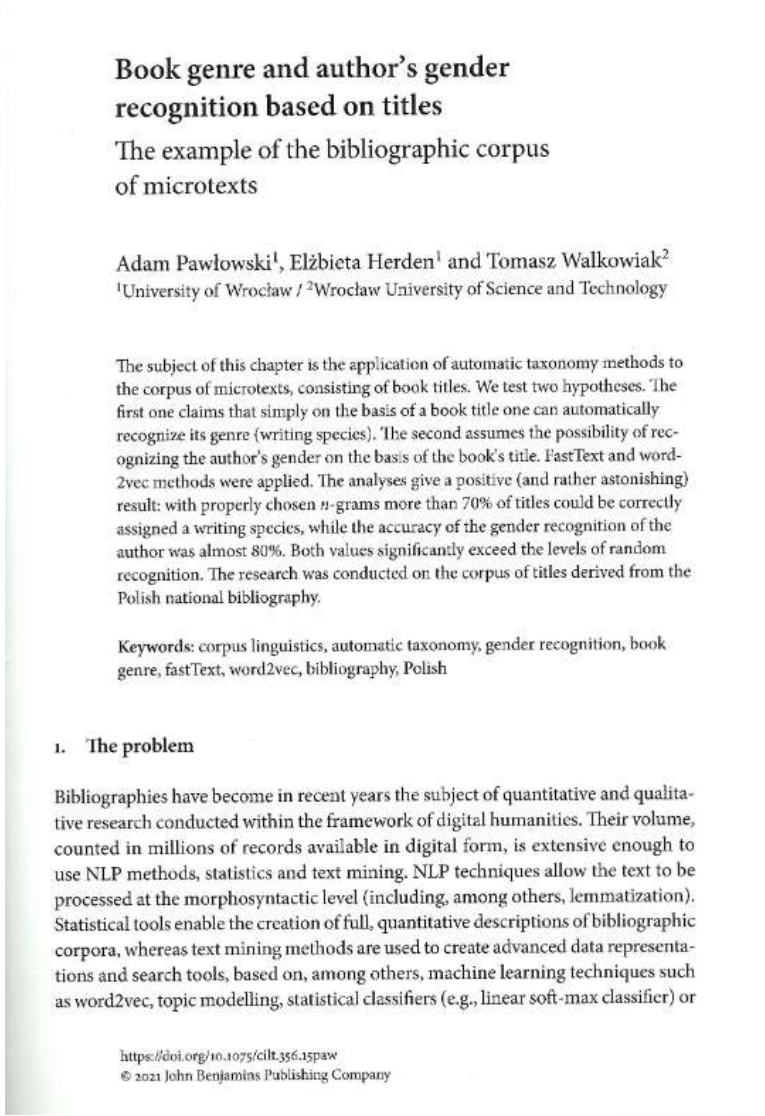

**Figure 1.** Histogram of title lengths in the bibliographic corpus

(Pawłowski, A., Topolski, K., & Herden, E., 2021, s. 244)

# BADANIA O CHARAKTERZE LINGWISTYCZNYM

## Adam Pawłowski, Elżbieta Herden, Tomasz Walkowiak (2021)

- Automatyczne rozpoznawanie płci autora i gatunków tekstów (m.in. naukowych i literackich) w mikrotekstach (opisach bibliograficznych)
- Źródło: „PB” – publikacje wydane w XX i XXI w. w j. polskim
- Ponad 1,8 mln rekordów
- Zastosowanie uczenia maszynowego, text mining, algorytmu AI „*fastText*” – narzędzia wytworzone w ramach konsorcjum CLARIN



**Book genre and author's gender recognition based on titles**

The example of the bibliographic corpus of microtexts

Adam Pawłowski[1], Elżbieta Herden[1] and Tomasz Walkowiak[2]
[1]University of Wrocław / [2]Wrocław University of Science and Technology

The subject of this chapter is the application of automatic taxonomy methods to the corpus of microtexts, consisting of book titles. We test two hypotheses. The first one claims that simply on the basis of a book title one can automatically recognize its genre (writing species). The second assumes the possibility of recognizing the author's gender on the basis of the book's title. FastText and word-2vec methods were applied. The analyses give a positive (and rather astonishing) result: with properly chosen *n*-grams more than 70% of titles could be correctly assigned a writing species, while the accuracy of the gender recognition of the author was almost 80%. Both values significantly exceed the levels of random recognition. The research was conducted on the corpus of titles derived from the Polish national bibliography.

**Keywords:** corpus linguistics, automatic taxonomy, gender recognition, book genre, fastText, word2vec, bibliography, Polish

1. The problem

Bibliographies have become in recent years the subject of quantitative and qualitative research conducted within the framework of digital humanities. Their volume, counted in millions of records available in digital form, is extensive enough to use NLP methods, statistics and text mining. NLP techniques allow the text to be processed at the morphosyntactic level (including, among others, lemmatization). Statistical tools enable the creation of full, quantitative descriptions of bibliographic corpora, whereas text mining methods are used to create advanced data representations and search tools, based on, among others, machine learning techniques such as word2vec, topic modelling, statistical classifiers (e.g., linear soft-max classifier) or

https://doi.org/10.1075/cilt.356.15paw

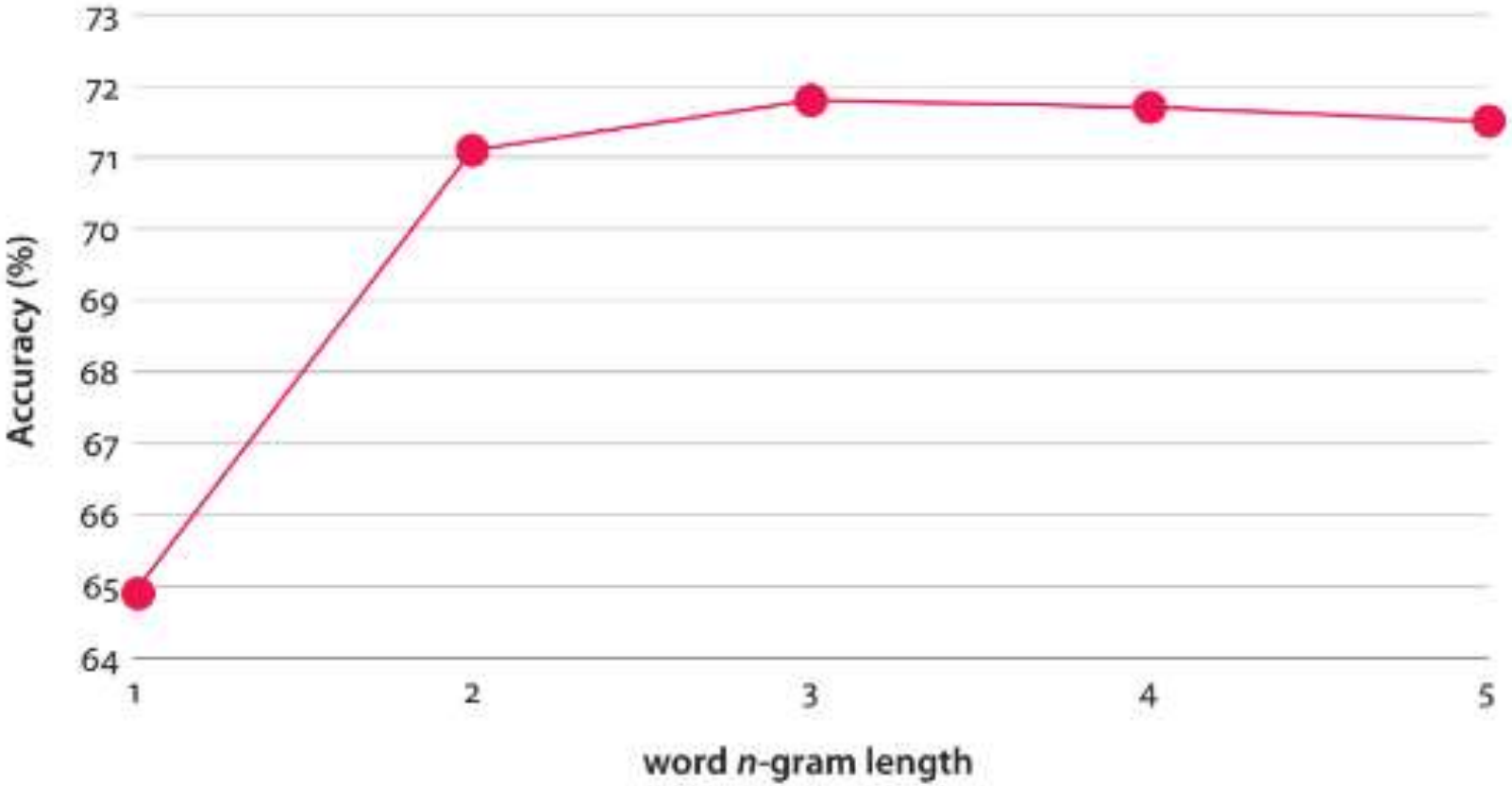



**Figure 1.** Accuracy of the literary genre attribution for the 'supervised fastText' method as a function of word *n*-grams length

Adam Pawłowski, Elżbieta Herden, Tomasz Walkowiak (2021)

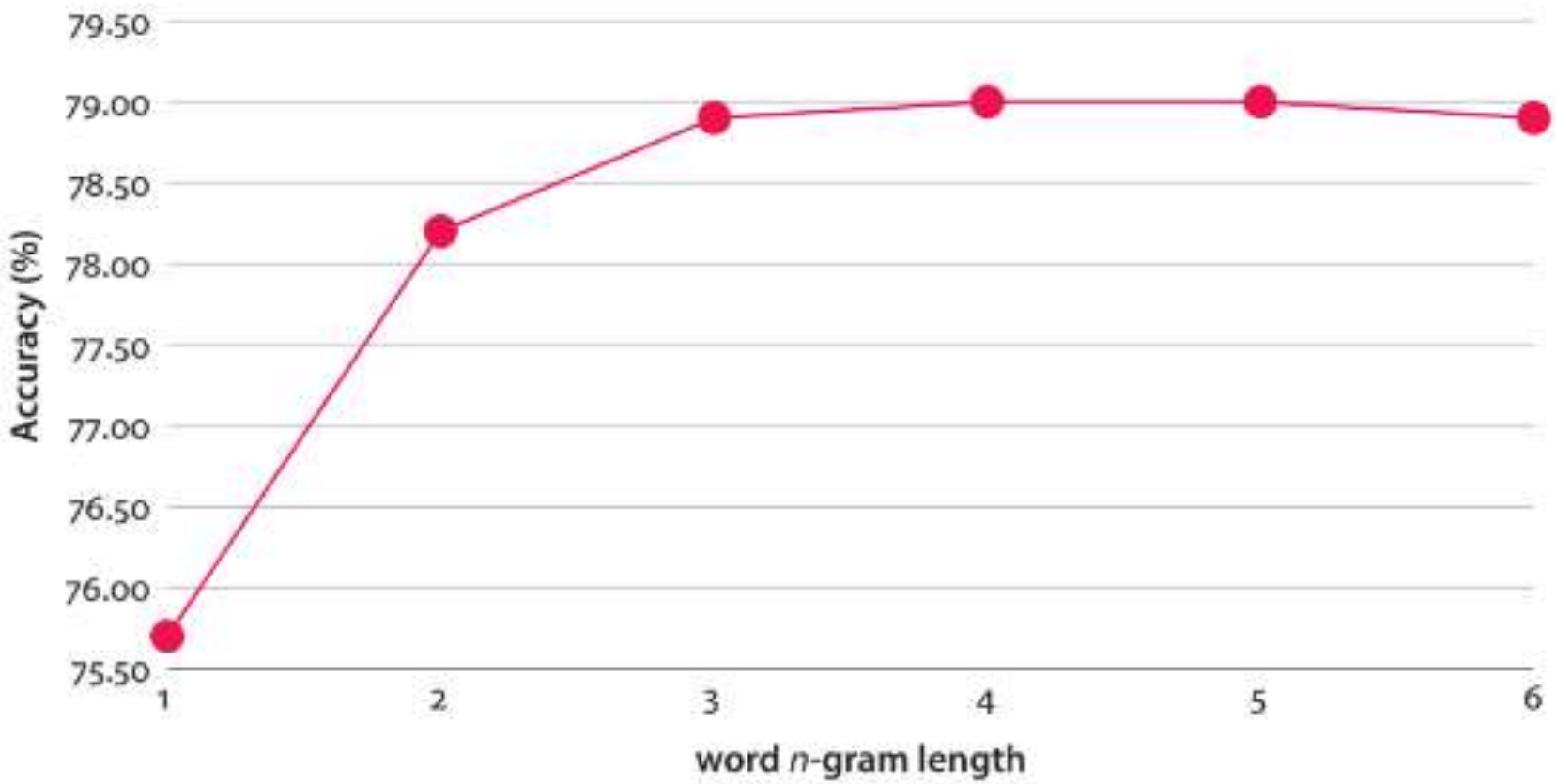


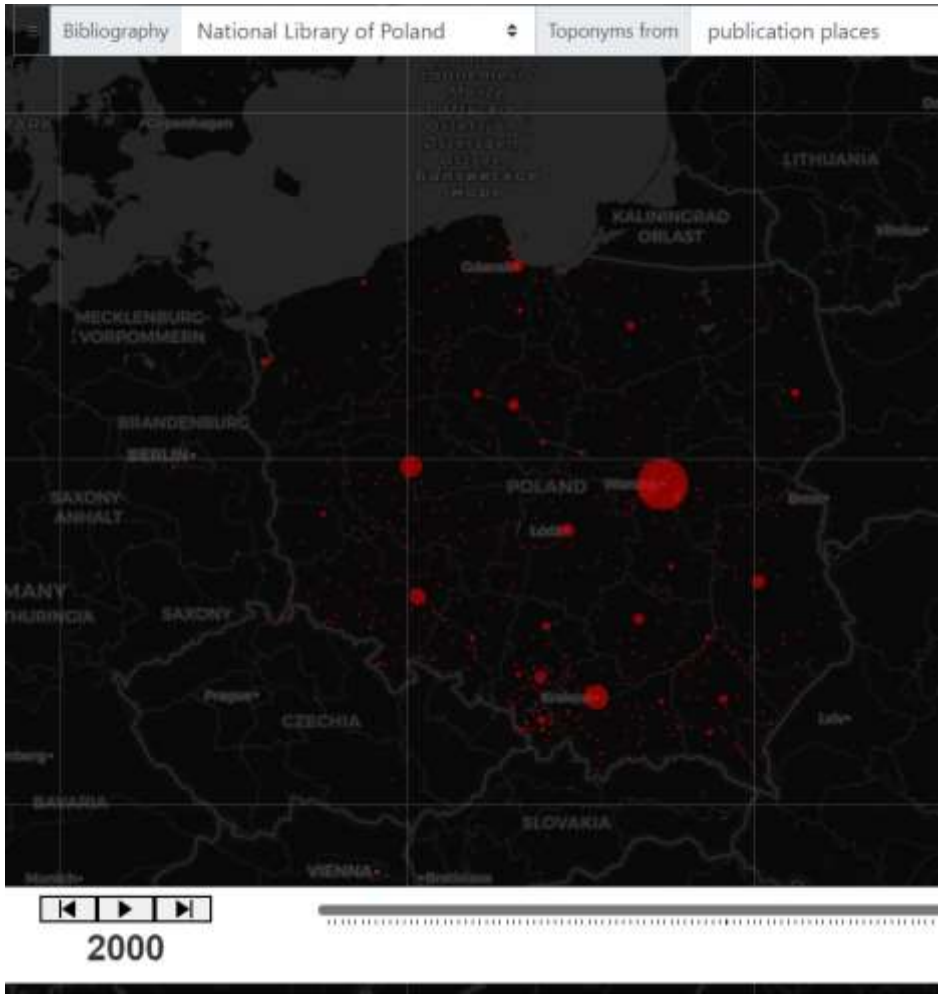

**Figure 2.** Accuracy of the gender attribution for 'supervised fastText' method in a function of word *n*-grams length

Adam Pawłowski, Elżbieta Herden, Tomasz Walkowiak (2021)

# BADANIA O CHARAKTERZE LINGWISTYCZNYM

**Adam Pawłowski, Tomasz Walkowiak (2021)**

- Źródło: Polska bieżąca bibliografia narodowa „Przewodnik Bibliograficzny”
- Metody lingwistyczne wspomagane eksploracją tekstu (*text mining*)
- Wrocław Bibliodata Website
  - Interaktywna mapa nazw miejscowych zawartych w rekordach katalogu Biblioteki Narodowej (1800–2019) i Polskiej Bibliografii Literackiej (druki zwarte od 1988)
  - https://bibgeos.clarin-pl.eu/maps.html
  - Interaktywna mapa nazw miejscowych wraz z wizualizacją kierunków geograficznych zawartych w tytułach i opisach publikacji ujętych w katalogach Biblioteki Narodowej (1860–2019)
  - https://bibgeos.clarin-pl.eu/distance_NER.html



# BADANIA DOTYCZĄCE METADANYCH

Związane z jakością metadanych (poprawność, spójność, wiarygodność, kompletność)

Wytwarzanie metod i narzędzi optymalizacji metadanychi ich wzbogacania

Cel: poprawa metadanych dla ułatwienia przeprowadzania badań na tego typu danych, tworzenia bardziej użytecznych danych dla badaczy

# BADANIA DOTYCZĄCE METADANYCH

## Evan Bryer et al. (2021)

- Metoda deduplikacji rekordów
- Wykorzystanie  metod i technologii uczenia maszynowego
- Podstawa: ponad 5 mln rekordów w formacie MARC 21 dla wydawnictw zwartych opublikowanych między XVI a XIX w.
- Źródło: WorldCat
- Cel: wypracowanie metod optymalizacji jakości metadanych na potrzeby badań nad produkcją wydawniczą

# BADANIA DOTYCZĄCE METADANYCH

## Andreas Lüschow i José Calvo Tello (2021)

- Metoda identyfikaji wybranych gatunków literackich w katalogach bibliotecznych
- Wykorzystanie  metod i technologii uczenia maszynowego, język Python i bibliotekę *scikitlearn*
- Źródło:
  - niemiecki katalog rozproszony Gemeinsamer Verbundkatalog (GVK) – opisy (ponad 740 000 rekordów)
  - Gemeinsame Normdatei – wykaz gatunków literackich (1319, z których 1150 zostało wykorzystanych w GVK)
- Cel: wypracowanie najbardziej efektywnego rozwiązania, które można zastosować w badaniach z wykorzystaniem dużych zasobów bibliograficznych

# BADANIA DOTYCZĄCE METADANYCH

## Róbert Péter et al. (2021)

- Wykorzystanie metod przetwarzania języka naturalnego i sztucznej inteligencji
- AVOBMAT – platforma do wizualizacji i analizy danych bibliograficznych i pełnych tekstów dokumentów
  - Pozwala na import danych z menedżera Zotero (CSV lub RDF)
  - Zaawansowane techniki wizualizacji i analizy dystrybucji wartości dla wybranych elementów metadanych
  - Obsługuje teksty w 52 językach (w tym j. polski)
- https://avobmat.hu

# DANE BIBLIOGRAFICZNE DANYMI BADAWCZYMI

## Dlaczego badać nowymi technologiami?

Poznanie prawidłowości – rozwiązanie problemu badawczego

Jest to możliwe

- ilość danych jest wystarczająca,
- dane są dostępne w postaci cyfrowej,
- pojawiły się metody automatycznego przetwarzania języka naturalnego (NLP),
- dane są wysokiej jakości, tworzone systematycznie a ich struktura jest w miarę jednolita

Wypracowanie spójnej metodyki pracy nad metadanymi (*metadata curation workflow*), którą inni badacze będą mogli zastosować

- → publicznie dostępna dokumentacja i pakiet narzędziowy

# DANE BIBLIOGRAFICZNE DANYMI BADAWCZYMI

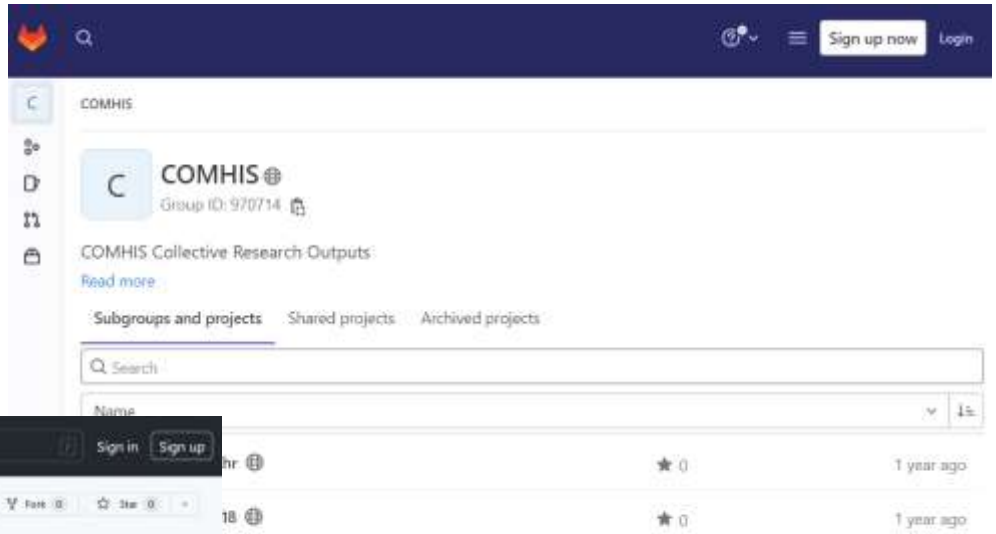
## Przy użyciu czego można badać?

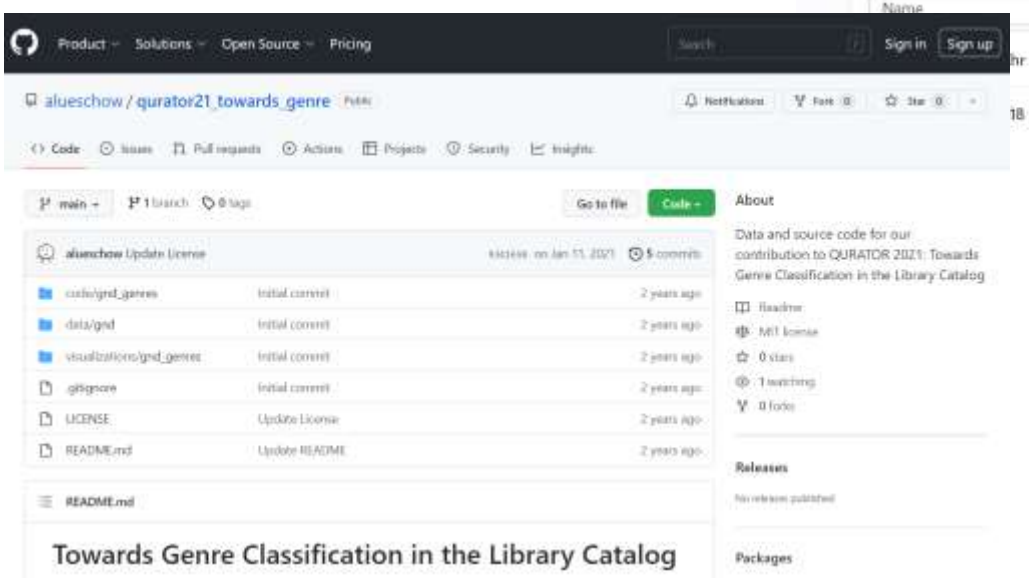
- Metody ilościowe
- Metody jakościowe
- Przetwarzanie języka naturalnego (NLP) za pomocą metod statystycznych i probabilistycznych
- Programowanie probablistyczne (*probablistic programming*)
- Metody i techniki automatycznej eksploracji danych (*data mining*) i tekstu (*text mining*)
- Uczenie maszynowe (*machine learning*) do automatycznej klasyfikacji elementów danych bibliograficznych
- Data Science (np. do odzwierciedlenie grafowej/sieciowej natury powiązań między dokumentami) – Python, R …

# INFRASTRUKTURA BADAWCZA

## Dokumentacja projektów

- GitLab
  - https://gitlab.com/COMHIS/
- GitHab
  - GitHub - alueschow/qurator21_towards_genre: Data and source code for our contribution to QURATOR 2021: Towards Genre Classification in the Library Catalog



# INFRASTRUKTURA BADAWCZA

Dostępne narzędzia:

- Wrocław Bibliodata Website – http://phc.uni.wroc.pl/wbw/
- CLARIN
- AVOBMAT (Analysis and Visualization of Bibliographic Metadata and Texts) - https://avobmat.hu
- WCRF Tagger – http://nlp.pwr.wroc.pl/redmine/projects/wcrft/wiki
- WCRFT2 – https://clarin-pl.eu/dspace/handle/11321/36

# BIBLIOMETRIA A BIBLIOGRAPHIC DATA SCIENCE

| Bibliometria | Bibliographic Data Science |
| --- | --- |
| • ocena stanu i rozwoju komunikacji piśmienniczej (głównie naukowej) z naciskiem na jej **produktywność i efektywność**<br>• pole badawcze **informacji naukowej** | • badanie produkcji wiedzy, główna uwaga jest skupiona na **uwarunkowaniach społecznych i kulturowych oraz kontekście historycznym**<br>• pole badawcze subdyscyplin HC<br>• metody i techniki analizy jakości metadanych realizowane z wykorzystaniem nowoczesnych technologii informacyjno-komunikacyjnych |

(M. Roszkowski, 2022, s. 22)

# JAKA PRZYSZŁOŚĆ DANYCH BIBLIOGRAFICZNYCH?

Nowe podejście do badań nad danymi bibliograficznymi

W jaki sposób uwzględnić je przy tworzeniu danych bibliograficznych?

Wypracowanie standardów dotyczących czyszczenia i poprawy jakości metadanych do badań

Współpraca środowiska

- DARIAH Bibliodata WG

# LITERATURA


Bryer, Evan; Rhujittawiwat, Theppatorn; Comandur, Samyu; Madrid, Vasco; Riley, Stephanie; Rose, John; Wilder, Colin. (2021). Analysis of Clustering Algorithms to Clean and Normalize Early Modern European Book Titles. ACM International Conference Proceeding Series, pp. 106-112. https://doi.org/10.1145/ 3451471.3451489

Lahti, L., Ilomäki, N., & Tolonen, M. (2015). A quantitative study of history in the english short-title catalogue (ESTC), 1470-1800. *LIBER Quarterly*, 25(2), pp. 87-116. https://doi.org/10.18352/lq.10112

Lahti, L., Mäkelä, Eetu; Tolonen, Mikko. (2020). Quantifying bias and uncertainty in historical data collections with probabilistic programming [online]. CEUR Workshop Proceedings, 2723, pp. 280-289. [dostęp: 17.01.2022]. Dostępny w WWW: http://ceur-ws.org/Vol-2723/short46.pdf

Lahti, L., Marjanen, J., Roivainen, H., & Tolonen, M. (2019). Bibliographic Data Science and the History of the Book (c. 1500–1800). *Cataloging & Classification Quarterly*, *57*(1), 5–23. https://doi.org/10.1080/01639374.2018.1543747

Pawłowski, A., & Walkowiak, T. (b.d.). *Analysis of Toponyms from the Polish National Bibliography*. https://CEUR-WS.org/Vol-2981/paper4.pdf

Pawłowski, A., & Walkowiak, T. (2020). Automatic Recognition of Gender and Genre in a Corpus of Microtexts. W W. Zamojski, J. Mazurkiewicz, J. Sugier, T. Walkowiak, & J. Kacprzyk (Red.), *Theory and Applications of Dependable Computer Systems* (s. 472–481). Springer International Publishing. https://doi.org/10.1007/978-3-030-48256-5_46

Pawłowski, A., & Herden, E. (2020). Przetwarzanie bibliografii metodami big data—Nowe możliwości czy stare ograniczenia? W *Big data w humanistyce i naukach społecznych* (s. 101–118). Wydawnictwo Naukowe i Edukacyjne SBP.

Pawłowski, A., Herden, E., & Walkowiak, T. (2021). Book genre and author's gender recognition based on titles. W *Language and Text: Data, models, information and applications*. John Benjamins Publishing Company.

Pawłowski, A., Topolski, K., & Herden, E. (2021). Quantitative analysis of bibliographic corpora: Statistical features, semantic profiles, word spectra. W A. Pawłowski, J. Mačutek, S. Embleton, & G. Mikros (Red.), *Language and text: Data, models, information and applications* (s. 239–256). John Benjamins Publishing Company. https://doi.org/10.1075/cilt.356.16paw

Péter, R., Szántó, Z., Seres, J., Bilicki, V., & Berend, G. (b.d.). *AVOBMAT: a digital toolkit for analysing and visualizing bibliographic metadata and texts*. 13.

Roszkowski, M. (2022). Bibliographic Data Science – konceptualizacja obszaru badawczego. *Przegląd Biblioteczny*, *90*(1), 5-26. https://doi.org/10.36702/pb.914

# DZIĘKUJĘ ZA UWAGĘ