Wikimedia versus traditional biographical encyclopedias. Overlaps, gaps, quality and future possibilities.

Daniel Baránek, Ph.D.
Institute of History, Czech Academy of Sciences

Lenka Křížová, Ph.D.
Institute of History, Czech Academy of Sciences

Abstract

This project aims to 1) analyze the current state of biographical entries on Wikipedia, Wikidata, and physical Central European dictionaries, 2) identify existing gaps and needs, and 3) propose solutions for Wikimedia projects and traditional dictionary production.

Introduction

Wikipedia has become the primary source of biographical information due to its extensive coverage. Traditional biographical dictionaries serve mostly only to cross-verify or find entries not yet covered by Wikipedia.

After years of competition, a mutual dependence between Wikipedia and traditional biographical dictionaries is evident. Wikipedia and Wikidata assist in creating traditional dictionaries, relying in turn on authority data generated by traditional producers of dictionaries like the Institute of History of the Czech Academy of Sciences. Wikipedia's voluntary content creation contrasts with the strictly systematic approach of traditional dictionaries.

This project seeks to:

- analyze the Central European, especially Czech, dictionary production,
- detect existing gaps and needs,
- propose solutions to enhance collaboration between Wikimedia and the creators of traditional biographical dictionaries,
- strengthen the content of Wikipedia and Wikidata.

The analysis will focus on biographical articles and entries of already deceased people who were born, lived, worked, or died in Czechia.

Date: July 1, 2024 – June 30, 2025.

Related work

Current scholarly studies on the production of traditional biographical dictionaries have so far taken into account the relationship to Wikipedia only to a very limited extent, both in the Czech (Sixta 2023) and European context. Biographical studies related to Wikimedia projects have mainly described the production of biographical content on Wikipedia (Graham 2015) or analyzed its content and shortcomings (e.g. Jemielniak 2016, Ribé 2021). Only a few studies (Carter 2019, Grote 2021) have delved more thoroughly into the relationship between traditional dictionary creation and Wikipedia, describing current problems in this area and proposing solutions.

Methods

The quantitative analysis will primarily identify national, political, gender, or other minority gaps in the content of Czech Wikipedia, Wikidata, and traditional biographical dictionaries. For instance, numerous people born and deceased in Czechia have an article in German but not in Czech. A comparison with traditional biographical dictionaries will assess their potential in filling these gaps.

The qualitative analysis will focus on individual articles on Czech Wikipedia and entities on Wikidata compared to traditional biographical dictionaries:

- The reference analysis hypothesizes that its biographical articles heavily rely on online accessible sources, overlooking scholarly literature from recent decades that is not accessible online due to copyright protection. An NLP model (HuggingFace) will be developed for the classification of the sources used in biographical articles. After identifying the temporal gaps of sources, we will suggest ways in which traditional biographical dictionaries can fill these gaps (e.g. by releasing licenses and transferring content onto Wikipedia).
- The content analysis of the biographical entries on Wikipedia and Wikidata will determine their completeness compared to traditional biographical dictionaries. For this purpose, several dictionaries will be probed (online/digitized/non-digitized, general/specialized, Czech/German-language, historical/contemporary, see References section). An OCR/HTR, segmentation (Kraken), and NLP model will be developed to extract basic data (date/place of birth/death, studies, occupation, works) from each entry. Based on the comparison of the datasets with Wikidata, solutions will be proposed to enhance Wikidata's quality by filling identified gaps.

Expected output

- International conference – platform for presenting the interim results of the analyses, formulating the ideas and opinions of the institutions producing biographical dictionaries and Wikimedia representatives, strengthening contacts, and finding common solutions to the current situation.
- Scientific publication – analysis and proposed solutions for Wikimedia and biographical dictionary developers.
- Machine learning models – HTR model, model for reference classification, and model for data extraction from biographical records for Wikidata contributors

Risks

Despite all efforts to set realistic targets, the following problems can arise:

- Datasets selected for analysis may prove to be too large. In such a case, a reasonable narrowing of the corpus will be done (e.g., discarding some of the dictionaries).
- There may be little interest from traditional dictionary producers to collaborate with Wikimedia. However, even such a situation would be an important finding in the analysis.

Community impact plan

- Naming specific shortcomings, gaps and needs of the Czech Wikipedia and Wikidata. The Czech Wikimedia branch will get a better idea of which areas

(gaps) need to be encouraged to be filled by launching editing contests.
- Establishing a quantitative and qualitative basis for the Czech Wikimedia office to negotiate with the creators of biographical dictionaries on mutual cooperation and data sharing.

Evaluation

The project can be evaluated as successful if:
- it detects gaps in Wikipedia and Wikidata content,
- it encourage the creators of traditional dictionaries to collaborate on Wikimedia projects (e.g., by providing identifiers or content).

Budget

Item	Calculation	USD
Personal costs: D. Baránek	12 months × 0.6 FTE * 2,900 USD	20,880
Personal costs: L. Křížová	12 months × 0.5 FTE * 2,900 USD	17,400
Personal costs: university student(s)	12 months × 0.5 FTE * 1,600 USD	9,600
Equipment: graphical card		920
Conference		1,200
Total		**50,000**

Prior contributions

The Institute of History of the CAS is a public research institution in the field of Czech and general history. It aims to transfer knowledge through various means, including online projects such as the Biographical Dictionary of the Czech Lands. In collaboration with the Czech Wikimedia branch, the IH has organized several workshops focused on Wikidata ([1] [2]).

Daniel Baránek has been a Wikipedian since 2006, active mainly on Czech Wikipedia and Wikidata. He has written several scholarly monographs and numerous articles on Jewish history. Recently, he has been focusing on the implementation of AI in historical research.

Lenka Křížová has been involved in the creation of the Biographical Dictionary of the Czech Lands as an author and editor since 2014 and later as a manager of its electronic version. She has authored a number of scholarly articles including those focused on biography.

References

General Biographical Dictionaries

- Josef Tomeš a kol., Český biografický slovník XX. století, d. 1–3, Praha 1999. (National Digital Library)
- Ottův slovník naučný. Ilustrovaná encyklopedie obecných vědomostí, d. 1–28, Praha 1888–1909. (Wikisource, National Digital Library)
- Biographisches Lexikon zur Geschichte der Böhmischen Länder, d. 1–4 (A–Štroner), München–Wien 1979. (Collegium Carolinum)
- Constantin Wurzbach, Biographisches Lexikon des Kaiserthums Oesterreich, enthaltend die Lebensskizzen derjenigen Personen, welche seit 1750 in den österreichischen Kronländern gelebt und gewirkt haben, d. 1–60, Wien 1856–1891; Register zu den Nachträgen, 1923. (Wikisource)

Specialized Biographical Dictionaries

- **Art:** Nová encyklopedie českého výtvarného umění, d. 1–2, ed. Anděla Horová, Praha 1995. (National Digital Library)

- **Theatre:** Eva Šormová a kol., Česká činohra 19. a začátku 20. století. Osobnosti 1–2, Praha 2015. (To be scanned)
- **Music:** Československý hudební slovník osob a institucí, d. 1–2, Praha 1963, 1965. (National Digital Library)
- **Literature:** Lexikon české literatury. Osobnosti, díla, instituce, d. 1; 2, sv. 1–2; 3, sv. 1–2; 4, sv. 1–2, Praha 1985–2008. (Institute of Czech Literature of the Czech Academy of Sciences)
- **Entrepreneurs:** Milan Myška a kol., Historická encyklopedie podnikatelů Čech, Moravy a Slezska do poloviny XX. století, d. 1–2, Ostrava 2003, 2008. (National Digital Library)
- **Archeology:** Karel Sklenář, Biografický slovník českých, moravských a slezských archeologů, Praha 2005. (National Digital Library)
- **Religion:** Český slovník bohovědný, ed. Josef Tumpach, Antonín Podlaha, d. 1–5 (A–Itálie), Praha 1912–1930. (National Digital Library)

Academic Biographistic Literature

- Maren Loren, Prezentace dějin na Wikipedii aneb touha po neměnnosti uprostřed konečné změny, *Dějiny – teorie – kritika* 11/1, 2014, pp. 122–144.
- Marie Makariusová, Slovenské biografické lexikony a Biografický slovník českých zemí v roce 2019, *Biografické štúdie* 42, 2019 s. 94-97.
- Václav Sixta, *Možnosti historické biografie. Teorie biografie a historická věda*, Praha 2023.
- Pamela Graham, "An Encyclopedia, not an Experiment in Democracy": Wikipedia Biographies, Authorship, and the Wikipedia Subject, *Biography* 38/2, 2015, pp. 222–224.
- Philip Carter, What is National Biography for? Dictionaries and Digital History, *True Biographies of Nations? The Cultural Journeys of Dictionaries of National Biography*, 2019, pp. 57–78.
- Dariusz Jemielniak, breaking the glass ceiling on Wikipedia, *Feminist Review* 113, 2016, pp. 103–108.
- Marc Miquel Ribé, Andreas Kaltenbrunner, Jeffrey M. Keefer, Bridging LGBT+ Content Gaps Across Wikipedia Language Editions, *The International Journal of Information, Diversity, & Inclusion*, 5/4 Special Issue, 2021 pp. 90–131.
- Jan Hodel, Wikipedia im Geschichtsunterricht, Frankfurt an Main 2020.
- Mathias Grote, Von Enzyklopädien zu Wikipedia und zurück?, *Aus Politik und Zeitgeschichte,* Bd. 71, Heft 3/4, 2021, S. 15–21.
- Thomas Wozniak (ed.), Wikipedia und Geschichtswissenschaft, Berlin 2015.