Wikimedia Research Fund 2024 - Stage II proposal

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.
3. Propose solutions for Wikimedia projects and traditional dictionary production.

Introduction

Academic Dictionaries and Wikimedia

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

This fact can be proven by comparing the Czech Wikipedia and the electronic version of the Biographical Dictionary of the Czech Lands (BDCL) produced by the Institute of History of the Czech Academy of Sciences. Considering only the entries published both on the Czech Wikipedia and in the BDCL, we can observe an average of 251 hits per day on the BDCL for such entries, while Wikipedia averages 23,161 hits per day, which is almost 100 times more. (The number of page views on Wikipedia was analyzed using the Wikimedia REST API.)

Additionally, only entries up to the letter H (more precisely Hi) are completed in the BDCL, which means approximately 6,000 entries completed out of a total of approximately 50,000 entries in the BDCL index. If we consider the entire index and the corresponding articles on the Czech Wikipedia, we find a daily average of 415 hits on the BDCL, but 171,649 on the Czech Wikipedia. This means that the impact of Wikipedia is more than 400 times greater. After years of jealous competition, it turns out that the dictionaries produced by scientific institutions cannot match the impact of Wikipedia.

However, a mutual dependence between Wikipedia and traditional biographical dictionaries is still evident. As editors of the BDCL, we know the important role Wikipedia and Wikidata play, at least in the initial stages of creating entries for biographical dictionaries. On the other hand, Wikipedia and even more so Wikidata rely on authority data generated by traditional producers of dictionaries like the Institute of History of the Czech Academy of Sciences and other research institutions.

Finally, the approaches to creating Wikipedia and traditional dictionaries can be complementary, as each has its advantages and disadvantages. While Wikipedia content is created more or less randomly by volunteers, research institutions take a strictly systematic approach.

Common Mission

As an academic institution, the Institute of History strives to approach history and its sources critically and not blindly reproduce contemporary discourse. Inherent in this is the uncovering of culturally conditioned disproportions and inequalities that may have seemed natural and acceptable to contemporaries but appear unjustifiable in hindsight.

This also applies to the creation of biographical dictionaries. The institutions producing these dictionaries are, to some extent, part of the "structures of power and privilege" (Redi, 2021, p. 5). Therefore, if they want to fulfill their scientific mission, they must constantly and purposefully reflect on whether they are producing the most historically faithful image of the past or merely reproducing stereotypes.

This moment is also crucial for Wikipedia because, due to the No Original Research (NOR) rule, it relies heavily on the output of research institutions. Recognizing the social responsibility of scientists, we are concerned about how the reconstructed image of the past is disseminated beyond professional circles. Because the immediate impact of our own production is limited, we must also be deeply concerned about how the image of history is disseminated through Wikimedia projects. Also for this reason, we care about "fostering conversations across the Wikimedia and academic communities" (Redi, 2021, p. 5).

Aims

Based on the situation described, 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 considered 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

Quantitative Analysis

The quantitative analysis will primarily identify language, cultural, gender, socio-economic, geographic, and other representation gaps in the content of Czech Wikipedia, Wikidata, and academic biographical dictionaries (for gap definitions, see Redi, 2021, pp. 21–24).

For instance, almost 400 people born and deceased in Czechia have an article in German but not in Czech. In the BDCL, for example, there are almost 3,000 entries that have no article in the Czech Wikipedia. The quantitative analysis is therefore intended to determine the potential of electronic and print academic dictionaries to enrich the content of Wikipedia and Wikidata and fill the mentioned gaps.

We chose the dictionaries for analysis in such a way that they cover different types of gaps. A list of selected academic dictionaries is given in the references.

Qualitative analysis

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

The content analysis of the biographical entries on Wikipedia and Wikidata will determine their completeness compared to traditional biographical dictionaries. Using the tools mentioned below, basic data (date/place of birth/death, studies, occupation, works) will be extracted from relevant entries on Wikipedia and Wikidata and from the entries in the traditional biographical dictionaries. Based on the comparison of the datasets, solutions will be proposed to enhance the quality of Wikipedia and Wikidata by filling identified gaps.

The reference analysis hypothesizes that biographical articles on the Czech Wikipedia heavily rely on online accessible sources, overlooking scholarly literature from recent decades that is not accessible online due to copyright protection. After identifying the temporal gaps of sources (cf. “recency gap”, Redi, 2021, pp. 22–23), we will suggest ways in which academic biographical dictionaries can fill these gaps (e.g., by releasing licenses and transferring content onto Wikipedia).

Tools and Workflow

1. The content of the selected printed dictionaries is obtained (digitized) using eScriptorium, Kraken, and segmentation and OCR/HTR models.

With this procedure, we have already digitized one volume of *Biographisches Lexikon zur Geschichte der Böhmischen Länder* as a trial. In doing so, we have created several effective models and already published the segmentation model: https://doi.org/10.5281/zenodo.10783346.

2. Based on the digitized content, an index of entries is compiled for each dictionary and used for quantitative analysis.

3. The digitized content is annotated by an NLP model using Wikidata identificators.

As a trial, we fine-tuned the mT5-small NLP model. In the first step, we constructed a training dataset from Wikidata (example) so that a model could be quickly created to recognize the initial structure of a dictionary entry and transform it into Wikidata format. We have published a beta version of the “biography2wikidata” model at HuggingFace: https://doi.org/10.57967/hf/1898.

For example, this model now annotates the introduction of a dictionary entry in this way:

Anschiringer, Anton, Publizist, * 1812 Wien, † 17. 12. 1873 Reichenberg (Liberec). Erzieher im Hause des Großindustriellen...

↓

{{WD|label|Anschiringer, Anton}}, {{WD|P106|Q6051619|Publizist}}, * {{WD|P569|1812}} {{WD|P19|Q1741|Wien}}, † {{WD|P570|1873-12-17|17. 12. 1873}} {{WD|P20|Q146351|Richenberg (Liberec)}}. Erzieher im Hause des Großindustriellen...

So already at this point, our text-to-text model can correctly process the dates to ISO format and assign the correct Wikidata property and item to common professions and birth/death places. Thus, this initial model already greatly accelerates the annotation of the first dataset, and further fine-tuning will accelerate the processing of the digitized content even more.

4. Each dictionary entry is assigned to a Wikidata item (Q109832485 in our example). The assignment is automated by several Wikidata Query Service queries checking the name, birth/death dates, or birth/death places.

5. At this moment, a comparison between the dictionary entry annotated as Wikidata-triples and actual Wikidata triples is made. In our case, we can see that the Wikidata item contains the same birth and death dates but no information about birth and death places. The missing information can be easily added to Wikidata and all the information can be sourced by the processed dictionary.

Anton Anschiringer	Biogr. Lex. zur Geschichte…	Wikidata (Q109832485)
birth date (P569)	1812	1812
birth place (P19)	Vienna	–
death date (P570)	1873-12-17	1873-12-17
death place (P20)	Reichenberg/ Liberec	–
occupation (P106)	opinion journalist	–

6. Similarly, the information about sources used in the selected biographical dictionaries is extracted from the dataset and from the article on the Czech Wikipedia (or other relevant language versions, mostly German). The comparison will confirm or refute the time gap hypothesis. If the gap hypothesis is confirmed, we will look for ways to remedy the situation. If the hypothesis is found false, our study will contribute to strengthening respect and trust for Wikipedia content, especially in professional circles.

The model for source classification has not been developed in the pilot project, but will be developed based on similar NLP techniques as in the case of Wikidata-triples extraction.

7. Using these methods, we therefore identify content gaps, create a dataset enabling to work on filling the gaps, and determine the quality and relevance of the sources used.

8. In the last step, the analysis will serve to make qualified strategic decisions: a) what forms the cooperation between Wikimedia and scientific institutions should take, b) what direction the creation of academic dictionaries and Wikipedia articles should take.

Expected output

1. International conference – a 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.

2. Scientific publication – analysis and proposed solutions for Wikimedia and biographical dictionary developers.

3. 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:

1. Datasets selected for analysis may prove to be too large. Our priority is to implement all steps of the workflow so that we can draw any analytical conclusions at all. Therefore, we perform the entire workflow on one dictionary

first before moving on to the next one. We will proceed from general dictionaries to more specialized ones, which of course include personalities which are less important from a general point of view. Therefore, if necessary, we omit the specialized ones from the corpus of analyzed dictionaries.

2. 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

1. Identifying 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.

2. The results of the analysis will indicate the direction in which it makes sense for the Czech Wikimedia branch and research institutions to cooperate. The analysis will establish a quantitative and qualitative basis for the Czech Wikimedia branch to negotiate with the creators of biographical dictionaries on mutual cooperation and data sharing.

3. The created HTR and NLP models will be published and can be used to enrich the content of Wikipedia and Wikidata in any language.

Evaluation

The project can be evaluated as successful if:

- it detects gaps in Wikipedia and Wikidata content,
- it encourages 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 (details)		1,200
Total		**50,000**

Response to reviewers and meta-reviewers

Reviewer 1 (EuUb)

The reviewer particularly questioned our proposal in Stage I, stating that it “lacks clarity on how it is relevant and impactful beyond Czech community” and recommended us to “examine how ML/GenAI could be used to close knowledge gaps”. We have therefore carried out a pilot project and elaborated our methodology to make it clear that the “bibliography2wikidata” model we have developed can be used for extracting Wikidata triples from Czech- and German-language biographical dictionaries without any fine-tuning, and for any language after fine-tuning.

Reviewer 2 (SFAK)

The reviewer in particular requested references for our claims regarding the relationship between biographical dictionaries and Wikipedia. Therefore, we tracked the traffic of the BDCL, which is most similar to Wikipedia in terms of the form of entries and the use of Mediawiki software, for three weeks. After comparing the number of page views on BDCL and the corresponding articles on the Czech Wikipedia, it turned out that the impact of

Wikipedia is 100-400 times higher than the impact of BDCL, depending on the set of entries taken into account.

The reviewer further supports our goal of analyzing gaps but objects to our goals "on improving collaboration and content creation" because they "are not research-focused". We agree with this objection in part, but by elaborating on the methodology section, we have made it clear that most of the project is devoted to developing machine learning tools (HTR and NLP models) to analyze dictionary production and to decect the gaps.

As a by-product of our research, several large datasets will be produced, and we consider it natural to use these datasets to enrich the content of Wikidata. Actually, this step does not require much effort and time thanks to tools like QuickStatements. In other words, in order to make comparisons and detect gaps, we need to extract Wikidata triples from the entries in traditional dictionaries. At this stage, it is just a small step to convert the Wikidata triples into a Wikidata database.

Moreover, it would be problematic to just publish the datasets produced and leave it to someone else to process them, since the content of datasets is subject to copyright protection, unlike extracted information that can be added to Wikidata. Of course, we could just publish the extracted Wikidata triples, but creating such an extracted dataset and publishing it would be about as much work as transferring the data directly to Wikidata.

With regard to strengthening cooperation, we agree that, in the strict sense of the word, this is not really research. On the other hand, we consider it to be an integral part of our scientific work that, if we carry out an analysis and detect a problem (gaps in this case), we try to find a qualified solution. And we already see this solution at this point in time in strengthening mutual cooperation. In the framework of our project, we only want to specify the forms of this cooperation. From our point of view as members of a research institution, sorting out scientific discourse at conferences and establishing collaborations is an integral part of scientific work. After all, like the enrichment of Wikidata content, these activities represent only a very small part of the workload of the proposed project.

In summary:

1. We have referenced our claims about the relationship between traditional dictionaries and Wikipedia in the introduction section.

2. The core of our project is to analyze the current state of the art and to detect gaps. Enriching the wikidata content within our project is just a small step, which we consider natural and not time consuming.

3. The point 2 applies also to the search for solutions to the current situation. Finding qualified solutions and fostering collaboration represent only a minor part of the work on a project, and we consider it an integral part of the scientific operation.

Reviewer 3 (xHp9)

The reviewer had comments only on the risks of our project. As suggested, we have elaborated on the prioritization in the risks section.

Metareview

The metareview recomended scaling back an focusing on a single dictionary, which corresponds with the comments of Reviewer 3. We have narrowed down the selection of dictionaries for analysis and elaborated on scaling back in the Risk section. However, we do not consider it necessary to limit the selection to a single dictionary. In the pilot project, we have gained an idea of the time required for each step and therefore consider our selection realistic.

The second requirement was to reference our claims, which corresponds to the comments of Reviewer 2. We have elaborated our claims and supported them with sources, especially in the Introduction section.

Grant Talk Page

No further comments appeared on the project discussion page by March 15, 2024.

References

General Biographical Dictionaries

- *recency gap:* Ottův slovník naučný. Ilustrovaná encyklopedie obecných vědomostí, d. 1–28, Praha 1888–1909. (Wikisource, National Digital Library)
- *language/national gap:* Biographisches Lexikon zur Geschichte der Böhmischen Länder, d. 1–4 (A–Štroner), München–Wien 1979. (Collegium Carolinum)
- *language/national gap:* 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

- *important topic gap (art):* Nová encyklopedie českého výtvarného umění, d. 1–2, ed. Anděla Horová, Praha 1995. (National Digital Library)
- *important topic (music)/recency gap:* Československý hudební slovník osob a institucí, d. 1–2, Praha 1963, 1965. (National Digital Library)
- *important topic (literature) gap:* 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)
- *socionomic status gap:* Milan Myška a kol., Historická encyklopedie podnikatelů Čech, Moravy a Slezska do poloviny XX. století, d. 1–2, Ostrava 2003, 2008. (National Digital Library)
- *cultural (religion) gap:* 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, pp. 94-97.
- Václav Sixta, *Možnosti historické biografie. Teorie biografie a historická věda*, Praha 2023.
- Miriam Redi, Martin Gerlach, Isaac Johnson, Jonathan Morgan, Leila Zia, A Taxonomy of Knowledge Gaps for Wikimedia Projects (Second Draft), 2021. arXiv:2008.12314.
- 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, pp. 15–21.
- Thomas Wozniak (ed.), Wikipedia und Geschichtswissenschaft, Berlin 2015.