

Mapping the Science of Interwar Czechoslovakia: Digital Reconstruction of Scientific Networks and Methodological Challenges

Filip Kersch ^{1,2}

¹Library of the Czech Academy of Sciences · ²Institute of Information Science and Librarianship,
Faculty of Arts, Charles University

1 Introduction

This poster presents a project aimed at reconstructing the information network of scientific publishing in Czechoslovakia between 1910 and 1939. The goal is to map the development of science during the period of the newly established state, when the domestic scientific community faced tensions between the need to demonstrate national independence, fulfil new state mandates, and maintain continuity with the Austro-Hungarian scientific legacy, while also establishing its profile within the context of international trends.

From the perspective of the history of science, this investigation is essential for understanding the influence of culture and politics on scientific discourse. However, traditional historiography is often limited to narratives favouring dominant figures and individual institutions (Míšková, Franc, and Kostlán 2018), or to overly general aggregate statistics. The presented project overcomes this deficit by integrating Digital Humanities methods with the theoretical framework of Information Science, viewing science as a complex process of externalisation and circulation of scientific knowledge.

Methodologically the project relies on the analysis of a large corpus of digitised scholarly periodicals, supplementing traditional approaches with the concepts of distant reading (Moretti 2000) and macroanalysis (Jockers 2013, pp. 3–10). The application of infometrics and specifically Social Network Analysis (SNA) allows for an examination of the scholarly communication not as a collection of isolated studies,

but as a macroscopic model of thematic, institutional, and personal relationships. By applying graph metrics such as density, centrality (closeness, betweenness) and community detection, the project identifies the structural interrelationships and trends within this information network.

The project focuses on three primary objectives:

- To map the publishing activities of scientific institutions, identify dominant topics, and track their transformations over time.
- To identify links between authors, locations, topics and institutions and to construct a complex network model of scientific communication.
- To analyse the influence of externalities (e.g., political and social changes, the economic crisis) on the stability and direction of this information ecosystem.

2 Methodology and Discussion of Digital Approaches

The main objective of this poster is a critical discussion of the selection and effectiveness of digital methods in the initial phase of processing a large corpus of historical texts. The research is based on the analysis of full-text data from scientific studies and scholarly articles published by scientific societies and other professional institutions in Czechoslovakia between 1910 and 1938. The text corpus is selected to ensure disciplinary diversity and institutional representativeness, with Czech digital libraries serving as the primary data source. Given that the primary inputs consist of automated OCR transcriptions of varying quality and high disciplinary diversity, the poster focuses on evaluating three fundamental methodological challenges:

- Discussing the possibilities of automated document segmentation in cases where digitised units lack detailed structural metadata at the article level. Comparing approaches based on typographic analysis with methods of semantic discontinuity detection (Reimers and Gurevych 2019; Kostelník, Beneš, and Hradiš 2025) for the precise identification of article boundaries.
- Comparing different approaches to Named Entity Recognition (NER) for the identification of persons (specifically authors of analysed publications), geographical names, and institutions. These data are subsequently integrated into a graph database (Neo4j), enabling the application of Social Network Analysis (SNA) methods (Painter, Daniels, and Jost 2019).
- Comparing traditional topic modeling methods (LDA) with modern neural approaches (Top2Vec, BERTopic) (Grootendorst 2022; Egger and Yu 2022). We focus on how neural models utilising word embeddings can handle historical scientific terminology and the role of Large Language Models (LLMs) in the subsequent interpretation and semantic labelling of topics.

3 Interactive Presentation of Outputs

The poster presents the current stage of the project based on a pilot data sample. Outputs will be available in the form of an interactive demonstration of various analytical results. Visitors will be able to explore the modelled network of relationships, visualise transformation of topics and discuss the limitations of automated OCR processing in historical research.

References

- Egger, Roman and Joanne Yu (May 6, 2022). “A Topic Modeling Comparison Between LDA, NMF, Top2Vec, and BERTopic to Demystify Twitter Posts”. In: *Frontiers in Sociology* 7, p. 886498. ISSN: 2297-7775. DOI: [10.3389/fsoc.2022.886498](https://doi.org/10.3389/fsoc.2022.886498). PMID: 35602001. URL: <https://pmc.ncbi.nlm.nih.gov/articles/PMC9120935/> (visited on 03/05/2026).
- Grootendorst, Maarten (2022). *BERTopic: Neural topic modeling with a class-based TF-IDF procedure*. arXiv: 2203.05794. URL: <https://arxiv.org/abs/2203.05794>.
- Jockers, Matthew Lee (2013). *Macroanalysis: Digital Methods and Literary History*. Urbana, Chicago, Springfield: University of Illinois Press. ISBN: 978-0252079078.
- Kostelník, Martin, Karel Beneš, and Michal Hradiš (2025). *TextBite: A Historical Czech Document Dataset for Logical Page Segmentation*. Version 1. DOI: [10.48550/ARXIV.2503.16664](https://doi.org/10.48550/ARXIV.2503.16664). URL: <https://arxiv.org/abs/2503.16664> (visited on 06/12/2026). Pre-published.
- Míšková, Alena, Martin Franc, and Antonín Kostlán (2018). *Bohemia Docta: The Historical Roots of Science and Scholarship in the Czech Lands*. Trans. by Melvyn Clarke and Hana Jirkalová. Historie. Praha: Academia. ISBN: 978-80-200-2639-2.
- Moretti, Franco (2000). *Conjectures on World Literature*. Vol. 1. London: New Left Review.
- Painter, Deryc T., Bryan C. Daniels, and Jürgen Jost (Sept. 2019). “Network Analysis for the Digital Humanities: Principles, Problems, Extensions”. In: *Isis* 110.3, pp. 538–554. ISSN: 0021-1753. DOI: [10.1086/705532](https://doi.org/10.1086/705532). URL: <https://www.journals.uchicago.edu/doi/full/10.1086/705532> (visited on 03/05/2026).
- Reimers, Nils and Iryna Gurevych (Nov. 2019). “Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks”. In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics. URL: <https://arxiv.org/abs/1908.10084>.

AI Usage Declaration:

The Gemini AI tool was used in the preparation of this abstract for linguistic refinement and stylistic revision. The author of the project remains fully responsible for the professional content, phrasing, and the final form of the text.