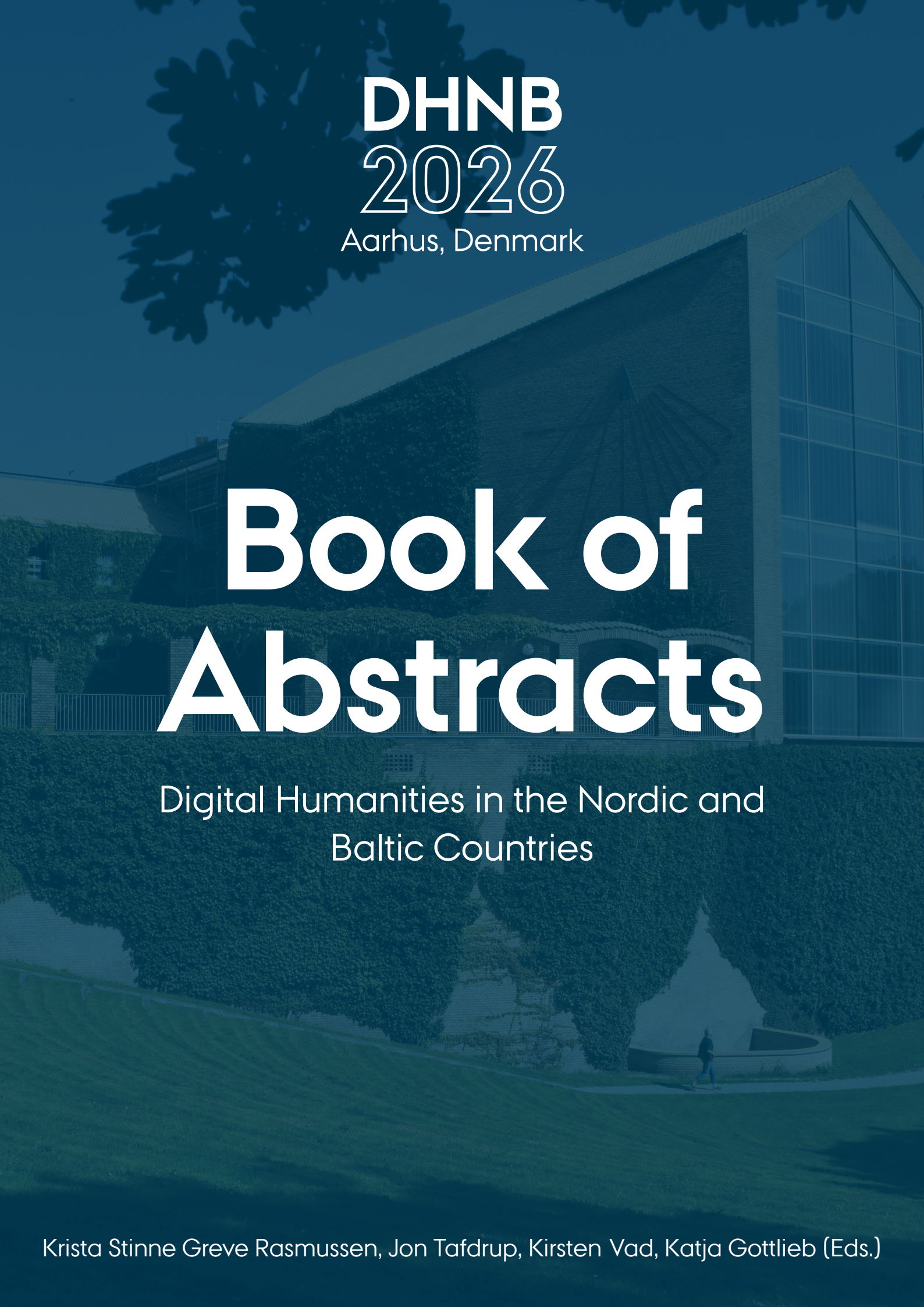

DHNB 2026

Aarhus, Denmark

Book of Abstracts

Digital Humanities in the Nordic and Baltic Countries

Krista Stinne Greve Rasmussen, Jon Tafdrup, Kirsten Vad, Katja Gottlieb (Eds.)

Book of Abstracts
&
Programme

The DHNB2026 conference, “Lost in Abundance: Encounters with the Non-Canonical”, aimed to foster fresh and critical approaches to uncovering overlooked literary trends, forgotten visual drifts, unconventional historical records, and other underrepresented cultural products. How might digital humanities guide us through this abundant yet underexplored landscape? What insights, challenges, and questions emerge when we re-direct our attention from canonical towards underexplored, neglected, or scattered data?

DHNB 2026 was organised by Center for Digital Textual Heritage and Center for Humanities Computing, Aarhus University.

Programme committee:

Katrine Laigaard Baunvig, Jon Tafdrup, Anda Baklane, Ernesta Kazakėnaitė, Pascale Feldkamp, Mari Väina, Eiríkur Smári Sigurðarson, Andres Karjus, Pille Runnel, Mads Rosendahl, Kristoffer Laigaard Nielbo, Krista Stinne Greve Rasmussen, Kirsten Vad, Fedja Borcak, Martin Eessalu, Johan Malmstedt, Jens Wendel-Hansen, Katja Gottlieb, Kim Steen Ravn, Paula Gheorgiade, Marta Kipke, Petra Hermankova, Zafar Hussein, Magnus Bender, Yevhen Kostiuk, Per Møldrup-Dalum, Mia Jacobsen, Kasper Fyhn, Giacomo Bilotti, Johan Heinsen, Camilla Bøgeskov, Peter Vahlstrup, Jeppe Büchert Netterström, Kristian Pindstrup, Natália Fedorova, Stephan Smuts, Lauritz Holm Petersen, Rie Schmidt Eriksen, Sara Kolding, Nicolas Legrand, Alie Lassche, Frida Hæstrup, Anne Agersnap

The conference took place March 9–13, 2026, at Aarhus University, Denmark.

Book of Abstracts edited by Krista Stinne Greve Rasmussen, Jon Tafdrup, Kirsten Vad, and Katja Gottlieb (Aarhus University). Cover design by Line Ejby Sørensen.

DOI for this publication: 10.5281/zenodo.18835192

Conference website: https://dhnb.eu/conferences/dhnb2026/

AARHUS UNIVERSITY

DHNB 2026

Digital Humanities in the Nordic and Baltic Countries

Programme

Aarhus, March 11–13 2026

Conference Programme Overview

WEDNESDAY, 11 MARCH 2026

08:30–10:30	Registration & Welcome Coffee			
10:30–12:15	**Plenary Session 1: Opening & Keynote — Mike Kestemont & Folgert Karsdorp** *(Stakladen 1423-111)*			
12:30–13:30	Lunch			
13:30–15:30	**Session 1A** *(Richard Mortensen Stuen 1422-122)*	**Session 1B** *(M1 1427-149)*	**Session 1C** *(M2 1427-246)*	**Panel 1** *(Preben Hornung Stuen 1422-132)*
	13:30–14:00 Quantifying Information Density in World-City Newspapers o… *Elena Fernandez Fernandez, Jana Meier, Simon Clematide*	*13:30–14:00* Levels of Canonicity: A Computational Analysis of Canon Fo… *Rie Schmidt Eriksen, Marta Kipke, Louise Brix Pilegaard Hansen, Yuri Bizzoni, Kristoffer L. Nielbo, Katrine L. Baunvig*	*13:30–13:50* The perception of song within the hymn material of N. F. S… *Anna Pouline Brogaard*	*13:30–15:30* Digitizing Medicine from Below: Researcher-Driven Digitiza… *Ylva Söderfeldt, Matts Lindström, Nils Hansson*
	14:00–14:30 News Circulation in Denmark-Norway (1770-1805): Preserving… *Alie Lassche, Rie Schmidt Eriksen, Kristoffer L. Nielbo, Katrine L. Baunvig*	*14:00–14:30* The Senses of Painting: Modeling Lexical and Semantic Patt… *Agata Hołobut, Szymon Pindur*	*13:50–14:20* Nordic Places of Worship (NordPoW) – an example of how to … *Stefan Gelfgren, Jakob Dahlbacka, Bo Ejstrud, Andreas Tjomsland*	
	14:30–15:00 Imperial War Experiences of the Finnish Soldiers in the Ba… *Aytac Yurukcu*	*14:30–14:50* Application of Multimodal AI in Classifying and Researchin… *Anda Baklāne, Valdis Saulespurēns, Alise Tifentale*	*14:20–14:40* Preaching During the COVID-19 Pandemic: A Quantitative Clo… *Emil Walther Bønding, Michael Mørch Thunbo, Anne Agersnap*	
	15:00–15:30 A World in Print: Introducing a Danish-Norwegian corpus of… *Johan Heinsen, Camilla Bøgeskov*	*14:50–15:10* Letters, Ledgers, and Lives: An Armenian Photographer's Ar… *Idil Cetin*	*14:40–15:10* From Flat Data to Deep History: A New Testament Corpus for… *Maciej Rapacz*	
			15:10–15:30 Modelling the Mythological Structure "Birth–Life–Death–Imm… *Olha Petrovych, Mari Väina*	
15:30–15:45	Break			
15:45–16:15	**1-minute Lightning Talks — Posters** *(Stakladen 1423-111)*			
16:15–18:00	**Poster Session**			
18:30–20:30	Welcome Reception			

THURSDAY, 12 MARCH 2026

08:30–9:45	**Plenary Session 2: Keynote — Katherine Bode** *(Stakladen 1423-111)*			
10:00–10:30	Coffee Break			
10:30–12:30	**Session 2A** *(Richard Mortensen Stuen 1422-122)*	**Session 2B** *(Preben Hornung Stuen 1422-132)*	**Session 2C** *(M1 1427-149)*	**Session 2D** *(M2 1427-246)*
	10:30–10:50 Imprints from the Women's Prison at Christianshavn 1870-1928 *Kamilla Matthiassen*	*10:30–10:50* Rule-based recognition of repetition *Ingerid Løyning Dale, Ranveig Kvinnsland*	*10:30–10:50* Consistency checking in a cloud of interlinked Cultural He… *Petri Leskinen, Annastiina Ahola, Heikki Rantala, Jouni Tuominen, Eero Hyvönen*	*10:30–10:50* Mapping Diversity in Norwegian Literature for Children and… *Lars G Bagøien Johnsen, Kristin Ørjasæter*
	10:50–11:20 Linking Heterogeneous Historical Sources: A Machine-Learni… *Tobias Kallehauge, Asbjørn Romvig Thomsen, Olivia Robinson, Anne Løkke, Barbara A. Revuelta-Eugercios*	*10:50–11:10* Lost in the Titles: Text-mining Metadata in the Digital Ed… *Kirsten Vad, Katrine Laigaard Baunvig*	*10:50–11:10* Between Print and Digital: Data Sharing and Community Form… *Sofia Papastamkou*	*10:50–11:10* Who is Sylvia? A comparison of plays by Albee and Rattigan *Maria Bekker-Nielsen Dunbar, Manex Agirrezabal Zabaleta*
	11:20–11:40 Migrate and visualise dispersed digitized data: A case stu… *Ole Jørgen Søndbø Abrahamsen*	*11:10–11:40* Leveraging Large Language Models for Lemmatization and Tra… *Lidia Pivovarova, Kati Kallio, Antti Kanner, Jakob Lindström, Eetu Mäkelä, Liina Saarlo, Kaarel Veskis, Mari Väina*	*11:10–11:30* ARCH-ON: A new ontological framework to describe archaeolo… *Pieterjan Deckers, Eero Hyvönen, Michael Lewis, Eljas Oksanen, Heikki Rantala, Jouni Tuominen*	*11:10–11:30* Molecules of a story: Community detection in narrative net… *Kasper Fyhn, Rebekah Baglini*
	11:40–12:00 From Transcription to Meaning: Digital Methods for Unlocki… *Jeppe Büchert Netterstrøm, Kristian Pindstrup*	*11:40–12:10* Computationally Identifying Recurrent Units in Finnic Oral… *Eetu Mäkelä, Kati Kallio, Mari Väina, Liina Saarlo, Jakob Lindström, Venla Sykäri, Antti Kanner, Maciej Janicki, Lidia Pivovarova, Kaarel Veskis*	*11:30–12:00* A Goldmine Unknown: Mapping and Visualizing the Saga Archi… *Åsa Warnqvist, Julia Beck*	*11:30–12:00* Using LLMs to uncover hidden patterns in the contestation … *David Zbíral, Zoltan Brys, Robert L. J. Shaw, Gideon Kotzé*
	12:00–12:30 Signals from the Field: A Study of Digital Practices and N… *Julia Kuhlin, Daniel Ihrmark, Koraljka Golub, Ahmad Kamal*		*12:00–12:30* Hidden Catastrophes: Mapping Disaster Narratives in Non-Ca… *Olena Koliasa*	*12:00–12:30* Trump, Catastrophe, and the Millennium: A Data-driven Stud… *Lauritz Holm Petersen*
12:30–13:30	Lunch			
13:30–15:30	**Session 3A** *(Richard Mortensen Stuen 1422-122)*	**Session 3B** *(Preben Hornung Stuen 1422-132)*	**Session 3C** *(M2 1427-246)*	**Panel 2 & 3** *(M1 1427-149)*
	13:30–13:50 From Old News to New Tools *Johan Heinsen, Matias Kokholm Appel*	*13:30–13:50* How About a Game of Sáhkku? How digitizing an Archive Empo… *Therese Foldvik, Ida Tolgensbakk*	*13:30–14:00* Non-Canonical Use of a Historical Dictionary: User Experie… *Sheryl McDonald, Tarrin Wills*	*13:30–14:30* Finding a needle in a haystack – user experiences with dig… *Mart Alaru, Pille Runnel, Agnes Aljas, Pille Pruulmann-Vengerfeldt, Kai Pata, Natali Ponetajev*

	13:50–14:10 Tracing Industrial Modernity in Global Historical Newspape… *Sophie-Marie Ertelt, Muhammad Okky Ibrohim, Andres Karjus*	*13:50–14:10* Appraising Archival Icebergs: Digitizing the Archive of th… *Christa Shusko*	*14:00–14:20* Icelandic eponyms: Detection and Representation in Lexicog… *Ellert Johannsson, Steinthor Steingrimsson*	*14:30–15:30* Uncovering Hidden Infrastructures for Digital Humanities i… *Mahendra Mahey, Pille Pruulmann-Vengerfeldt, Hans Dam Christensen, Berndt Clavier, Rikke Lie Halberg, Paula Bray*
	14:10–14:30 Reanimating early Danish periodical journals (1740-1770) a… *Maria Nørby Pedersen*	*14:10–14:30* Fragments, bias, lost voices and measuring silence in the … *Catherine Beck*	*14:20–14:40* The effect of translatorial signals in the source language… *Manex Agirrezabal, Seán Vrieland, Iliana Kandzha*	
	14:30–14:50 Exploring AI-Supported Qualitative Data Analysis *Daniel Andersson*	*14:30–14:50* The urge to digitise: Medieval materials in museum and arc… *Olga Zabalueva, Polina Ignatova*	*14:40–15:10* Неком Дрина тече десно, неком Дрина лијево тече (For some,… *Sasha Rudan Kelbert, Lazar Kovacevic, Sinisa Rudan*	
	14:50–15:10 Using Large Language Models for searching explainable rela… *Annastiina Ahola, Petri Leskinen, Heikki Rantala, Jouni Tuominen, Eero Hyvönen*	*14:50–15:20* A Best-Practice Pipeline for Nuanced, Reproducible Biblioq… *Eetu Mäkelä, Thea Lindquist*		
15:30–16:00	Coffee Break			
16:00–18:00	**Session 4A** *(Richard Mortensen Stuen 1422-122)*	**Session 4B** *(Preben Hornung Stuen 1422-132)*	**Session 4C** *(M1 1427-149)*	**Session 4D** *(M2 1427-246)*
	16:00–16:30 E-motion: Binary systems versus fluid identities *Onur Kilic, Evelina Liliequist, Coppelie Cocq, Karin Danielsson*	*16:00–16:30* Lost in Structure: Graph-Based Technologies for Digital Sc… *Sebastian Enns, Andreas Kuczera*	*16:00–16:30* "The Multimodal Television Generation": Experimental Synch… *Johan Malmstedt*	*16:00–16:30* Detecting Climate Delay Discourses in Danish Parliamentary… *Camilla Buur Küseler, Florian Meier*
	16:30–17:00 Navigating Abundance: A Platform for AI-Augmented, Hybrid … *Sasha Rudan Kelbert, Eugenia Kelbert Rudan*	*16:30–17:00* From Manuscript to Scripture: The Sanctification of N.F.S.… *Jon Tafdrup, Katrine Laigaard Baunvig*	*16:30–17:00* Tracing Terrorism in Television: Toward a Methodology of L… *Daniel Brodén, Johan Malmstedt, Mats Fridlund*	*16:30–17:00* Decoding Diplomacy: A Large Language Model Approach to Emo… *Sasha Nielsen*
	17:00–17:30 Computational Phonosemantics at Scale: Measuring Sound-to-… *Szymon Pindur*	*17:00–17:30* What Happens to Style After Success? A Multidimensional An… *Marc Barcelos*	*17:00–17:20* Scene-Anchored Analysis of Affective "Stickiness" in Nordi… *Aida Gholami*	*17:00–17:20* Tracing sentiment in the political discourse on homosexual… *Anna Maria Ramm*
	17:30–18:00 Modeling Textual Emotions in Literary Fiction *Kirstine Nielsen Degn, Alexander Conroy, Xiaoyuan Jiang, Ali Al-Laith, Jens Bjerring-Hansen, Tanya Karoli Christensen, Ingo Zettler, Daniel Hershcovich*		*17:20–17:40* Framing Digital Transformation: Media Discourses on Digita… *Coppélie Cocq, Stefan Gelfgren, Rebecka Weegar*	*17:20–17:40* From Low-Code to Open Code: Reflecting on a Developing Wor… *Daniel Ihrmark, Hanna Carlsson, Fredrik Hanell*
				17:40–18:00 Carniolan Provincial Assembly: Corpus Improvements and Enh…

				Ajda Pretnar Žagar, Kristina Pahor de Maiti Tekavčič
18:00–19:00	**DHNB Annual General Meeting** *(Stakladen 1423-111)*			
19:30–23:00	Conference Dinner			

FRIDAY, 13 MARCH 2026

	Session 5A *(Richard Mortensen Stuen 1422-122)*	Session 5B *(Preben Hornung Stuen 1422-132)*	DT 1 *(M1 1427-149)*	DT 2 *(M2 1427-246)*
08:30–10:30	*08:30–09:00* Retouched but Not Restored: Using a Generative Model to En… *Lina Samuelsson, Daniel Brodén, David Alfter, Johan Malmstedt*	*08:30–08:50* Scattered throughout (Digital) Libraries: The Case of Old … *Ernesta Kazakėnaitė*	*08:30–09:00* The odd one out: Conceptualizing and digitizing ephemeral … *Holger Berg*	*08:30–09:00* Hot Topics in the Parliament: A Topic Landscape Analysis w… *Anna Ristilä*
	09:00–09:20 Printing Possibility: Scaling inconspicuous labour adverti… *Sofus Landor Dam*	*08:50–09:10* Lifting Local Treasures: Building Digital Access to Paper … *Mia Gulvad Jørgensen, Andy Stauder*	*09:00–09:30* How digital methods are taught: A mixed-method analysis of… *Ahmad Kamal, Daniel Ihrmark, Patrick Gavin*	*09:00–09:30* A Community in Motion: Mapping Japanese American Migration… *Saara Kekki*
	09:20–09:50 "Were He Not Called Sigurðr" – Personal Name Case Studies … *Elisabeth Maria Magin*	*09:10–09:40* Close reading, automation and cultural memory. Experiments… *Alexander Conroy, Kirstine Nielsen Degn, Jens Bjerring-Hansen, Ali Al-Laith, Daniel Hershcovich, Matthew Wilkens*	*09:30–10:00* Assessing AI Recognition of Text and Symbols in Early Mode… *Thomas Holgersson, Daniel Ihrmark, Henrik Svensson, Jonas Svensson, Ahmad Kamal*	*09:30–10:00* Developing an Automatic Scansion Model for Early Modern Da… *Niels Nykrog, Manex Agirrezabal*
	09:50–10:10 Named Entity Recognition in the Historical Meeting Protoco… *Siim Orasmaa, Kadri Muischnek, Sofia Kriuchkova*	*09:40–10:10* Annotation Fever! An Interdisciplinary Experiment Explorin… *Klara Källström, Thobias Fäldt, Bernard Geoghegan, Mats Fridlund*		
	10:10–10:30 Women in Business on the Rhine in the 18th Century: Compar… *Lauri Matias Heinonen*	*10:10–10:30* Where the Machine Looks Away *Teodora Crisan-Matcaboja*		
10:30–11:00	Coffee Break			
11:15–12:30	**Plenary Session 3: Keynote — Bolette Sandford Pedersen** *(Stakladen 1423-111)*			
12:45–13:30	Lunch			
13:45–15:00	**Plenary: Closing Session** *(Stakladen 1423-111)*			
15:00–15:30	Coffee & Goodbye			

Index of Contributions

WEDNESDAY, 11 MARCH 2026

Session 1A *13:30–15:30*

Session 1B *13:30–15:10*

Session 1C *13:30–15:30*

Panel 1 *13:30–15:30*

Poster Session *16:15–18:00*

THURSDAY, 12 MARCH 2026

Session 2A *10:30–12:30*

Session 2B *10:30–12:10*

FRIDAY, 13 MARCH 2026

DT 1 *08:30–10:00*

DT 2 *08:30–10:00*

Session 5A *08:30–10:30*

Session 5B *08:30–10:30*

WEDNESDAY, 11 MARCH 2026

Session 1A — 13:30–15:30

13:30–14:00 *LONG PAPER* **[1]**

Quantifying Information Density in World-City Newspapers of Record and a Suburban Newspaper (1999–2018)

Elena Fernandez Fernandez, Jana Meier, Simon Clematide
University of Zurich, Switzerland

Keywords: *Information Density, World City Theory, Discourse Analysis, Social Acceleration*

Economists and geographers describe world cities not only as major urban centres but also as influential hubs of economic, cultural, geographic, and technological innovation. Building on this idea, several scholars argue that world cities constitute autonomous ecosystems increasingly detached from their national contexts. This paper investigates whether such an ecosystem can be detected through patterns of information behaviour, thereby challenging geography as the only explanatory factor. We analyse a multilingual dataset of eight newspapers headquartered in a sample of Western world cities: seven newspapers of record (The Times, The Irish Times, Le Figaro, Die Welt, NZZ, La Stampa, El País) and one suburban newspaper: Chicago Daily Herald—covering the years 1999–2018. We treat the suburban newspaper as an exploratory study case, to be expanded in future work. Then, in dialogue with Franzosi's concept of narrative units, we extract and quantify unique Subject–Verb–Object (SVO) triplets per token per year as a proxy for information density. The analysis reveals a consistent increase in unique SVO triplets across the seven newspapers of record, suggesting a convergent pattern of information density among these outlets. By contrast, the Chicago Daily Herald shows a declining trend despite its geographic location within a world city. This divergence challenges the assumption that location alone explains informational homogeneity. The results also extend research on the Theory of Social Acceleration, indicating that newspapers of record reflect increasing informational activity, whereas our suburban outlet does not. We discuss possible confounding factors such as editorial policies, content scope, and media materiality.

14:00–14:30 *LONG PAPER* **[2]**

News Circulation in Denmark-Norway (1770-1805): Preserving Content Diversity through Topic-Vector Representations

Alie Lassche[1], Rie Schmidt Eriksen[1], Kristoffer L. Nielbo[1], Katrine L. Baunvig[2]
[1] *Center for Humanities Computing, Aarhus University, Denmark*
[2] *Center for Grundtvig Studies, Aarhus University, Denmark*

Keywords: *historical newspapers, Denmark-Norway, embeddings, BERTopic, information spread, event detection*

Abstract

We study how news circulated across the composite monarchy of Denmark–Norway by combining large-scale newspaper data with a representation that preserves content heterogeneity. Using a sample of a re-digitized corpus (8 titles, 17,600 editions; 0.75M articles, 1770–1805), we first filter for news (vs. advertisements) with a four-class classifier trained on a 1,700-item gold sample. We then derive topic distributions per article from embeddings using BERTopic, and aggregate these to topic-vector representations per newspaper and date. With cosine similarity within and between newspapers, we trace diffusion pathways, speeds, and center-periphery dynamics, situating patterns against historical constraints (temporary press freedom under Struensee; subsequent censorship; postal/infrastructural factors). The approach contributes (1) a scalable method for representing multi-entry historical newspapers without losing semantic diversity and (2) new empirical insight into geographically fractured, heterogeneous information ecosystems in late-eighteenth-century Denmark–Norway.

1. Introduction and Related Work

This paper addresses two intertwined challenges: how to study news dissemination through newspapers in the historical conglomerate state of Denmark-Norway, and how to represent newspaper content computationally without flattening its complexity. These challenges are not separate – the historical question of how information circulated through a geographically dispersed society with few urban centers demands a methodological approach that preserves rather than reduces the semantic diversity of newspaper content.

Emerging in the early seventeenth century, newspapers gradually evolved into a mass medium that, by the late eighteenth century, delivered political news, advertisements, and local announcements to a broad public (Weber 2006; Pettegree 2014). Beyond conveying information, the press fostered shared discussion within the public sphere and contributed to the formation of imagined national communities (Habermas 1989; Anderson 2016). In the composite monarchy of Denmark–Norway, newspapers expanded from serving Copenhagen's political and mercantile elite to reaching readers across the provinces (Kjærgaard 1989). Their contents diversified accordingly: reports of wars, natural disasters, and royal ceremonies appeared alongside advertisements from farmers, servants, book shops, and tradespeople (Søllinge and Thomsen 1988).

Drawing on a large-scale corpus of digitized eighteenth- and early nineteenth-century newspapers, we investigate how news moved between urban and rural communities, between Denmark and Norway, and what these patterns reveal about information dynamics in a society with a small center and a large periphery. Studying this circulation at scale poses a key methodological problem: how to represent newspaper content in a way that captures semantic and topical diversity rather than extracting only dominant signals.

Most computational studies of Danish historical newspapers have focused on case studies of specific topics (Agersnap et al. 2025; Heinsen and Birkemose 2023; Baunvig 2021, 2023). As the scale of the corpus increases, the analytical methods must scale accordingly. Techniques such as document embeddings and topic modeling make large-scale analysis possible, but they risk obscuring the heterogeneity of content within individual newspapers or time periods. We therefore seek a representation that is computationally manageable for thousands of newspaper articles while preserving their internal variation.

Other studies have represented newspaper content as a single vector. For instance, in research on Danish newspapers from the COVID-19 pandemic and on twentieth-century Dutch newspapers, a representation was created from the linguistic content of the front page, based on the assumption that modern newspapers highlight current and pressing events there (Wevers, Kostkan, and Nielbo 2021; Nielbo et al. 2021). However, because the style and layout of historical newspapers differ entirely from their modern counterparts, this approach is not suitable in our context. In a study on early modern Dutch chronicles – a text in which one day can be described in multiple entries with various topics – each day was represented by the semantic embedding with the highest centrality among that day's entries, with

the uncertainty of that choice expressed as a standard deviation (Lassche et al. 2022). This method, however, falls short here, as we do not wish to discard the less central articles in our newspapers.

In this paper, we combine semantic embeddings with topic modeling to create topic distributions for each article, which we then aggregate to represent newspapers over time. Rather than reducing a week of newspapers to a single mean embedding or a handful of keywords, we maintain a vector representation that captures which topics appear and in what proportions. This allows us to trace not just whether and at what speed and by which routes news spread from center to periphery, but how the topical composition of newspapers varied across regions and evolved over time.

This methodological innovation enables us to examine a question that has remained difficult to answer empirically: how did local newspapers participate in the circulation of knowledge? Drawing on scholarship on media systems and information flow, we know that censorship, literacy rates, postal routes, and infrastructural and geological constraints shaped news dissemination in Denmark-Norway (Hoyer 1968; Drotner 2011). Peripheral newspapers occupied a complex position: although dependent on urban sources for national and international news, they also created space for local concerns and perspectives to emerge in print (Hoyer 1968; Søllinge, Thomsen, and Danske). By preserving the diversity of content within each newspaper, our method allows us to trace this duality: to see both how central news reached the periphery and how peripheral newspapers constructed their own distinctive information landscapes.

In doing so, we contribute to our knowledge about nineteenth-century information dynamics. Rather than treating the ‘public sphere’ as a uniform entity (Habermas, Lennox, and Lennox 1974), we show how collective knowledge formation was geographically fractured and semantically heterogeneous – a mosaic of overlapping but distinct information ecosystems across the Denmark-Norway conglomerate state.

2. Corpus

We draw on a sample from the Danish historical newspaper dataset created at Aalborg University (Heinsen and Bøgeskov 2025). Due to the poor OCR quality of the newspapers, they were re-digitized via the Transkribus platform, for which custom layout and text recognition models were trained on the collection. In addition to transcription, newspaper editions were segmented into individual articles. The result is a dataset comprising nearly five million articles from 28 different newspapers, published during and after the composite state of Denmark–Norway between 1666 and 1850.

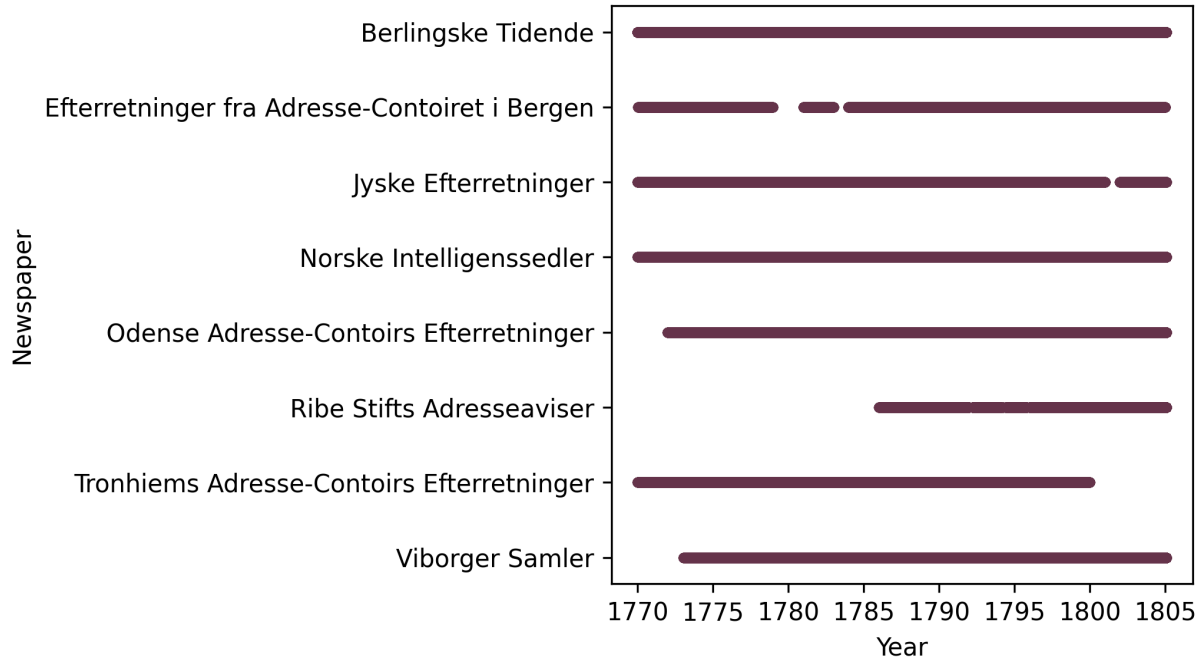

Figure 1: Print run per newspaper in our sample.

Newspaper	Editions	Articles	Words
Berlingske Tidende	3,941	250,306	40,648,088
Efterretninger fra Adresse-Contoiret i Bergen	1,618	59,052	3,987,917
Jyske Efterretninger	2,259	70,281	7,953,273
Norske Intelligenssedler	1,828	104,812	7,431,554
Odense Adresse-Contoirs Efterretninger	3,266	120,104	11,387,719
Ribe Stifts Adresseaviser	1,371	48,135	3,763,929
Tronhiems Adresse-Contoirs Efterretninger	1,553	31,313	2,784,509
Viborger Samler	1,792	56,998	6,158,078
total	17,628	741,001	84,115,067

Table 1

Descriptive statistics on the newspapers in the dataset.

Since the newspapers vary in their years of publication, we select for this study a subset of titles and a time span in which multiple newspapers overlap. Our sample consists of eight newspapers, encompassing more than 17,000 editions and 0.75 million articles, covering the years 1770 to 1805 (see Table 1). Figure 1 shows the individual print run of the newspapers in our sample. The chosen period includes newspapers from both Denmark and Norway and coincides with several key moments that shaped the circulation of news. Most notably, it encompasses the brief era of press freedom under J. F. Struensee, the physician to King Christian VII and *de facto* ruler of Denmark, whose 1770 decree abolished censorship. The ensuing explosion of print lasted until his execution in 1772, after which restrictions were reimposed (Langen and Stjernfelt 2022). The chosen time frame also captures major national and international developments, including the gradual abolition of the *stavnsbånd* (serfdom)

between 1788 and 1800, the ban on the Atlantic slave trade (1792, effective 1803), and the American, French, and Haitian revolutions.

3. Methods

To characterize information spread in our corpus, we propose the methodological pipeline visualized in Figure 2.

3.1. Filter news articles from sample

We focus exclusively on news articles, excluding advertisements. Using a gold sample of 1,700 manually annotated articles from an earlier study on the historical newspaper corpus (Lassche et al. 2026), we train a four-class Logistic Regression model to predict the categories *National news*, *International news*, *Advertisement*, and *Miscellaneous*. The model uses embeddings from the Old_News_Segmentation_SBERT_V0.1 model,[1] which achieved the best performance in our benchmark experiments. Figure 3 shows the number of *National* and *International news* articles per newspaper and month. The resulting dataset contains 327,532 news articles.

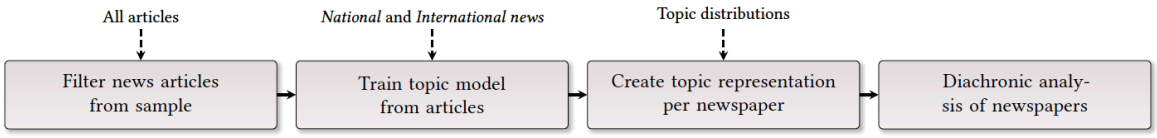

Figure 2: Four-stage pipeline for characterizing information spread in newspapers.

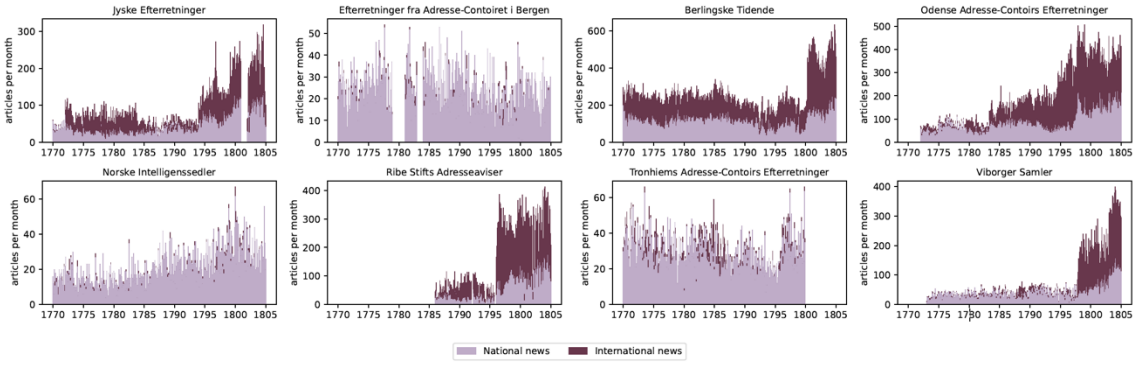

Figure 3: Articles per newspaper, colored by category.

3.2. Train topic model from articles

We have a semantic embedding for each article, but to create a representation of all articles within a single newspaper – or within a specific week or month – we need a method that captures the semantic content of a group of articles without simply averaging their embeddings.[2] To achieve this, we use the semantic embeddings as input for a topic model, which allows us to represent the dominant themes in a newspaper's content. This approach is similar to earlier studies (Lassche et al. 2022) in which we

[1] https://huggingface.co/JohanHeinsen/Old_News_Segmentation_SBERT_V0.1.
[2] Averaging embeddings can flatten semantic distinctions and obscure topical diversity.

implemented the Top2Vec package (Angelov 2020), but since Top2Vec supports only a limited range of embedding models, we turned to the BERTopic package (Grootendorst 2022), which works in a similar way, but is more flexible towards the use of precomputed embeddings. The algorithm consists of the following steps:

1. Reducing semantic embeddings to two dimensions with UMAP (McInnes et al. 2018);[3]
2. Finding clusters of articles with HDBSCAN (McInnes, Healy, and Astels 2017);[4]
3. Converting the documents of each cluster into sparse bag-of-words representations using CountVectorizer;
4. Identifying the most representative words per topic by applying class-based TF–IDF (c-TF/IDF);
5. Optionally merging or splitting topics based on semantic similarity to improve coherence.

Our topic model consists of 506 topics, which we reduce to 142 topics by manually merging similar topics.

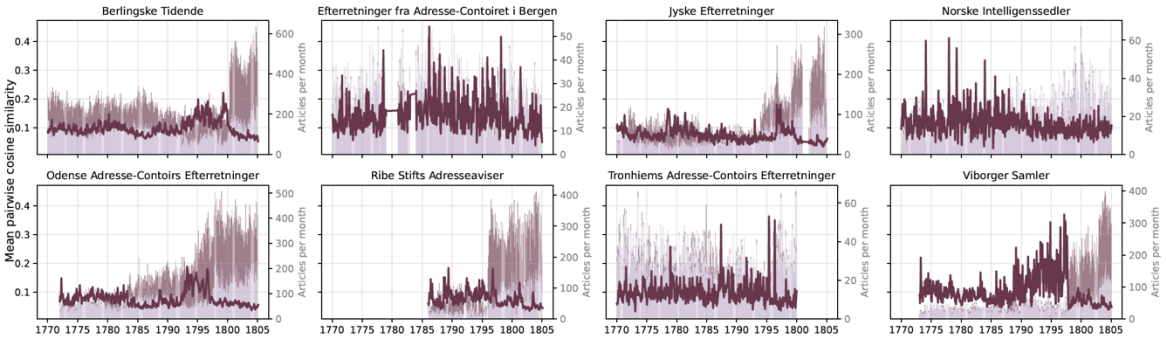

Figure 4: Intra-group similarity across time, using the mean pairwise cosine similarity of articles in a rolling window (s = 14, $step$ = 1) over the days for each newspaper title. The signal is smoothened with a Gaussian filter (σ = 8). The bar plot in the background and the right y-axis show the number of articles for that newspaper. Note that newspapers with International News articles show a stable low intra-group similarity.

3.3. Create topic representation per newspaper

We aggregate the topic distributions of all articles within each newspaper by date, counting how many articles are most associated with each topic. This yields a vector of topic weights – one per newspaper and date – which is normalized to allow comparison over time.

3.4. Diachronic analysis of newspapers

We measure topic fluctuations and within-newspaper (intra) and between-newspaper (inter) cosine similarity to track how topical composition changes over time and how newspapers influence one another within the broader information landscape.

[3] We ran several experiments with varying parameter settings and found the best results with $n\_neighbors$ = 50, $n\_components$ = 2, min_dist = 0.0.

[4] We ran several experiments with varying parameter settings and found the best results with $min\_cluster\_size$ = 40.

4. Results

Figure 4 shows the intra-group similarity across time for each newspaper title. The figures show that the newspapers that contain an *International News* section, show a more stable pattern of low intra-group similarity, suggesting a stable and predictable editorial formula, where international news sections impose a consistent topical structure across editions. In our final presentation, we will furthermore focus on the role of peripheral newspapers in the Danish-Norwegian mediascape, their relationship to the main Copenhagen-based newspaper *Berlingske Tidende*, and the consolidation of the Danish public sphere. We will finish with discussing whether the use of topic-vector representations of Danish newspapers is a fruitful way to study the news circulation in Denmark-Norway.

References

Agersnap, Anne, Katrine Baunvig, Line Wittoff Schmidt, Rie Schmidt Eriksen, Emil Walther Bønding, Thomas Husted Kirkegaard, and Lea Wierød Borčak. 2025. “The Human Touch: Leveraging HITL for Quantitative Close Reading of Historical Corpora.” In DHNB 2025 Book of Abstracts, 11–13. Tartu.

Anderson, Benedict R. O’G. 2016. Imagined Communities: Reflections on the Origin and Spread of Nationalism. Revised edition. London New York: Verso.

Angelov, Dimo. 2020. “Top2Vec: Distributed Representations of Topics.” arXiv:2008.09470 [cs, stat].

Baunvig, Katrine Frøkjær. 2021. “Fictional Realities of Modernity: The Fantastic Life of Demi-Goddess Dana in the Emerging Nation State of Denmark.” In Mythology and Nation Building: N.F.S. Grundtvig and His Contemporaries. Aarhus University Press.

Baunvig, Katrine Frøkjær. 2023. “‘Each of Our Springs Has Lost Its Miraculous Power’: The Range of a Religious Hotspot – A Distant Reading of Lourdes Representations in Denmark 1858–1914.” Numen 70, no. 1: 43–69.

Drotner, Kirsten. 2011. Mediehistorier. 1. udgave. Frederiksberg: Samfundslitteratur.

Grootendorst, Maarten. 2022. BERTopic: Neural Topic Modeling with a Class-Based TF-IDF Procedure, arXiv:2203.05794.

Habermas, Jürgen. 1989. The structural transformation of the public sphere: an inquiry into a category of bourgeois society. MIT Press.

Habermas, Jürgen, Sara Lennox, and Frank Lennox. 1974. “The Public Sphere: An Encyclopedia Article (1964).” New German Critique, no. 3, 49–55.

Heinsen, Johan, and Anders Dyrborg Birkemose. 2023. “Efterlyst. Identitet, tvang og mobilitet, 1750-1850.” Temp - tidsskrift for historie 14, no. 27: 24–53.

Heinsen, Johan, and Camilla Bøgeskov. 2025. A World in Print: Introducing a Danish-Norwegian Corpus of Historical Newspapers, arXiv:2509.02356.

Hoyer, Svennik. 1968. “The Political Economy of the Norwegian Press.” Scandinavian Political Studies 3.

Kjærgaard, Thorkild. 1989. “The Rise of Press and Public Opinion in Eighteenth-century Denmark—Norway.” Scandinavian Journal of History.

Langen, Ulrik, and Frederik Stjernfelt. 2022. The World’s First Full Press Freedom: The Radical Experiment of Denmark-Norway 1770–1773. De Gruyter.

Lassche, Alie, Pascale Feldkamp, Yuri Bizzoni, Katrine Baunvig, Kristoffer Nielbo, and Johan Heinsen. 2026. “Evaluating Embedding Models on Danish Historical Newspapers: A Corpus and Benchmark Resource.” In LREC2026.

Lassche, Alie, Jan Kostkan, Kristoffer Nielbo, Folgert Karsdorp, and Kristoffer Nielbo. 2022. “Chronicling Crises: Event Detection in Early Modern Chronicles from the Low Countries.” In Proceedings of the Computational Humanities Research Conference 2022. Antwerp: CEUR.

McInnes, Leland, John Healy, and Steve Astels. 2017. “Hdbscan: Hierarchical Density Based Clustering.” The Journal of Open Source Software 2 (11): 205.

McInnes, Leland, John Healy, Nathaniel Saul, and Lukas Grossberger. 2018. “UMAP: Uniform Manifold Approximation and Projection.” The Journal of Open Source Software 3 (29): 861.

Nielbo, Kristoffer L., Frida Haestrup, Kenneth C. Enevoldsen, Peter B. Vahlstrup, Rebekah B. Baglini, and Andreas Roepstorff. 2021. When No News Is Bad News – Detection of Negative Events from News Media Content, arXiv:2102.06505.

Pettegree, Andrew. 2014. The Invention of News: How the World Came to Know about Itself. New Haven; London: Yale University Press.

Søllinge, Jette D., and Niels Thomsen. 1988. De danske aviser 1634-1989 (I). Odense: Dagspressens Fond.

Weber, Johannes. 2006. “Strassburg, 1605: The Origins of the Newspaper in Europe.” German History 24 (3): 387–412.

Wevers, Melvin, Jan Kostkan, and Kristoffer Nielbo. 2021. “Event Flow - How Events Shaped the Flow of the News, 1950-1995.” In Computational Humanities Research Conference, 62–76. Amsterdam.

14:30–15:00 *LONG PAPER* **[3]**

Imperial War Experiences of the Finnish Soldiers in the Balkan Encounters and Their Letters to Finnish Newspapers and Home in 1877-1878.

Aytac Yurukcu

University of Eastern Finland, Finland

Keywords: *Digital history, emotional history, Finnish soldiers, collective memory, newspaper letters, imperial loyalty, encounters, national identity, correspondence networks.*

This paper examines a largely overlooked corpus of Finnish soldiers’ letters from the Russo-Ottoman War of 1877–1878 to explore how imperial loyalty and emerging national identity were negotiated through emotional and mediated communication. While scholarship on nationalism and empire has extensively theorized identity formation (Anderson 1991; Hobsbawm 1992; Brubaker 1996), and Finnish historiography has addressed the war primarily at institutional and diplomatic levels (Laitila 2001; Parppei 2021), the affective experiences and communicative practices of Finnish rank-and-file soldiers have remained marginal. This study addresses that gap by analysing soldiers’ correspondence as non-canonical sources that illuminate the emotional and translocal dimensions of imperial warfare.

The paper draws on a corpus of twenty-two published anonymous letters appearing in ten Finnish newspapers and approximately thirty private letters sent to family members during the Balkan campaign. These materials are examined using a mixed methodology that combines computational text analysis with close reading. Digital methods include keyword extraction and topic modeling to identify recurring emotional and ideological motifs; sentiment analysis to trace shifts between enthusiasm, ambivalence, and distress; and network analysis mapping the circulation of letters across newspapers, places of origin, and thematic clusters rather than interpersonal social ties. These distant-reading techniques are used heuristically and are complemented by close readings that situate individual voices within the broader emotional regime of nineteenth-century European warfare.

The analysis shows that Finnish soldiers often reproduced Russian imperial rhetoric emphasizing Christian liberation and humanitarian duty, yet their letters simultaneously reveal homesickness, moral uncertainty, and a growing sense of Finnish distinctiveness. When published in newspapers, private emotions were reframed through editorial practices, producing a layered discourse in which personal

affect intersected with imperial ideology and national interpretation. Censorship, mediation, and circulation thus shaped not only what was said, but how war was emotionally understood within Finland.

By foregrounding soldiers' letters as translocal, emotionally charged, and digitally recoverable sources, this paper contributes to debates in digital humanities, memory studies, and translocal history. It demonstrates how computational analysis of abundant digitized newspapers can recover marginalized voices while also requiring careful methodological reflection on scale, interpretation, and affect. Ultimately, the study repositions Finnish soldiers' correspondence as a crucial archive for understanding the emotional infrastructures of imperial war and the negotiated formation of national consciousness. By engaging critically with digital abundance and methodological scale, the paper reflects on the epistemological stakes of computational approaches to historical emotion. It argues that digital methods do not replace close reading, but make visible patterns of emotional circulation otherwise inaccessible.

Ultimately, this study situates Finnish soldiers' wartime writings within a multi-layered framework that combines digital corpus analysis, emotional discourse studies, and translocal history. It reveals how the Balkans, imagined as a site of moral duty and Christian brotherhood, became a mirror through which Finnish soldiers—and, by extension, Finnish society—negotiated questions of belonging, loyalty, and identity in the imperial context. (Doynov,1978; Drake; 1909; Kansanaho, 1965; Snellman and Kalleinen, 2022). In revisiting these forgotten letters, we encounter not only the soldiers' voices but also the affective infrastructures that sustained imperial wars and shaped national imaginaries. The paper thus contributes to a richer understanding of the emotional and communicative dimensions of nineteenth-century warfare, highlighting how digital humanities can recover and reinterpret the non-canonical traces of imperial pasts.

References

Anderson, Benedict. 1991. Imagined Communities: Reflections on the Origin and Spread of Nationalism. Verso.

Brubaker, Rogers. 1996. Nationalism Reframed: Nationhood and National Question in New Europe. Cambridge University Press.

Doynov, Stefan. Bŭlgarskata obshtestvenost i Rusko-turskata osvoboditelna voĭna [The Bulgarian public and the Russo-Turkish war of liberation (1877–1878)], Sofia, 1978.

Drake. L.L. 1909. "Iz Zhizni Russkikh Voysk v Finlyandii v 70-kh i 80-kh godakh". [From the Life of Russian Troops in Finland in the 70s and 80s]. Russkaya Starina. Vol. 137, January-March.

Hobsbawm, Eric J. E, 1992. Nations and nationalism since 1780: programme, myth, reality, 2nd ed., Cambridge.

Kansanaho, Erkki. 1965. "Punainen Risti ja Turkin sota: Suomen Punaisen Ristin syntyvaiheita", Historiallinen Aikakauskirja, 63, 193–207.

Kokko, Heikki. 2021. "Temporalization of Experiencing: First-Hand Experience of the Nation in Mid-Nineteenth Century Finland", In: Kivimäki, V., Suodenjoki, S., Vahtikari, T. (eds) Lived Nation as the History of Experiences and Emotions

in Finland, 1800-2000. Palgrave Studies in the History of Experience. Palgrave Macmillan, 109-133.

Laitila, Teuvo. 2001. Soldier, Structure and the Other. Social Relations and Cultural Categorization in the Memoirs of Finnish Guardsmen Taking Part in the Russo-Turkish War 1877-78. PhD. Thesis, Helsinki. Univ. 2001.

Parppei, Kati. "This Battle Started Long Before Our Days …' The Historical and Political Context of the Russo-Turkish War in Russian Popular Publications, 1877–78". Nationalities Papers, vol. 49, 1, 2021, 162-179.

Snellman Alex and Kalleinen Kristiina. 2022. "Introduction: Finland in Imperial Context", Journal of Finnish Studies 25, 2, 143-153.

15:00–15:30 *LONG PAPER* **[4]**

A World in Print: Introducing a Danish-Norwegian corpus of historical newspapers

Johan Heinsen, Camilla Bøgeskov

Department of Politics and Society, Aalborg University, Denmark

Keywords: *Historical newspapers, Transkribus, BERT-model, ENO, dataset reconstruction*

This Data Descriptor introduces the dataset Enevældens Nyheder Online (News during Absolutism Online). The Enevældens Nyheder Online (ENO) dataset provides a reconstruction of the contents of

major newspapers in Denmark and Norway during the period of Absolutism (1660–1849). The dataset contains approx. 474 million words, created using neural networks designed to process digitised microfilm versions of Danish newspapers as well as a smaller selection of Norwegian publications that were all hitherto illegible for computers. The contributions details this process and its results, including a way to derive standalone texts from the editions, and the accompanying BERT-model trained on a beta-version of the dataset.

Session 1B — 13:30–15:10

13:30–14:00 **Levels of Canonicity: A Computational Analysis of Canon Formation in 19th Century Danish Painting**
Rie Schmidt Eriksen, Marta Kipke, Louise Brix Pilegaard Hansen, Yuri Bizzoni, Kristoffer L. Nielbo, Katrine L. Baunvig

14:00–14:30 **The Senses of Painting: Modeling Lexical and Semantic Patterns in Audio Descriptions of Paintings Across Art Movements**
Agata Hołobut, Szymon Pindur

14:30–14:50 **Application of Multimodal AI in Classifying and Researching Zenta Dzividzinska's Photo Archive**
Anda Baklāne, Valdis Saulespurēns, Alise Tifentale

14:50–15:10 **Letters, Ledgers, and Lives: An Armenian Photographer's Archive and the Ethics of Digital Curation**
Idil Cetin

13:30–14:00 *LONG PAPER* **[5]**

Levels of Canonicity: A Computational Analysis of Canon Formation in 19th Century Danish Painting

Rie Schmidt Eriksen[1], Marta Kipke[1], Louise Brix Pilegaard Hansen[1], Yuri Bizzoni[1], Kristoffer L. Nielbo[1], Katrine L. Baunvig[2]

[1] *Center for Humanities Computing, Aarhus University, Denmark*
[2] *Center for Grundtvig Studies, Aarhus University, Aarhus, Denmark*

Keywords: *Image Classification, Network Analysis, Machine Learning, Cultural Heritage*

In this paper we investigate the dynamics of canon formation in Danish Golden Age paintings using techniques of visual embeddings and network analysis. Canon formation is an intricate phenomenon, especially in that time frame and medium. Thus, this case study deeply intertwines art-historical theory and historical contexts with computational methods and statistical measures. Our dataset consists of 1,656 digitized paintings from SMK (The National Gallery of Denmark, SMK), annotated with metadata including artist, date, and canonical status. We address the context and the challenges of this dataset, before we move on to the methodology of visual embedding and network construction. By correlating network measures - centrality, community position, and cluster specificity - with canonical status, we assess whether canonized artworks occupy distinctive structural positions within the corpus. Situating the analysis within art-historical and museological debates, we highlight how both 19th century curatorial practices (e.g., N.L. Høyen's nationalist selection criteria) and 21st-century digitization initiatives (notably SMK Open) contribute to ongoing processes of canon formation. Our findings enable us to discuss on a broader scale the different reasons a painting becomes canon and the distinct types of canonicity that emerge within the corpus, while also considering how processes of digitization and curation in contemporary museum practice actively shape and reinforce these hierarchies.

14:00–14:30 *LONG PAPER* **[6]**

The Senses of Painting: Modeling Lexical and Semantic Patterns in Audio Descriptions of Paintings Across Art Movements

Agata Hołobut, Szymon Pindur
Jagiellonian University in Krakow, Poland

Keywords: *audio description, semantic features, part of speech, linear models, art movements*

Introduction

“The verbal-vocal description of visual or audiovisual content for visually impaired audiences” (Hirvonen and Wiklund 2021), audio description is a service offered to blind and partially sighted viewers to allow them a richer experience of films, television productions, exhibitions, and live events.

Museum audio description, which is of specific concern to this paper, makes visual art accessible to non-sighted audiences. The practice is governed by professional guidelines, which help select and sequence information. Prototypically, title, author, style, and physical dimensions of the artwork are provided first, followed by the identification of the subject matter and a detailed description of the piece, according to a pre-determined scanning path. Verbal description of an artwork can be delivered live to accompany touch tours and other multisensory experiences offered by museums, or pre-recorded as an audio guide or voice recording posted on museum website.

Practiced in its current shape since the 1980s, it has received little academic attention. Most research on museum AD adopts a qualitative approach and pertains to the much-debated professional recommendations: scholars disseminate good practices and present particular applications in the form of case studies. Many contributions from within art history, museum studies, and pedagogy promote the technique, discuss the usefulness of AD for therapeutic and educational purposes, and provide art-philosophical reflections on its value (e.g., Pawłowska & Sowińska Heim 2016, 2019; Jerzakowska 2016).

Some scholars focus on the inadequacy of extant guidelines for re-evocation of aesthetic experience that visual art offers (Neves 2012; Więckowski 2014). They assume that visual art is by nature indeterminate and ambivalent, both features requiring subjective resolution on the part of the experiencer-describer. Since aesthetic visual experience cannot be reduced to a matter-of-fact description of its visual components, they either draw parallels with the ancient “word-painting” tradition of literary ekphrasis and recommend “non-objective” AD styles (Neves 2010, Soler et al. 2023, Bartolini 2023, Vizcaíno 2023, Reviers& Hanoulle 2023) or they reflect on the insufficiency of verbal resources as such, advocating experimental multisensory experiences to complement AD (Neves 2012, Randaccio 2020).

Crucially for the current project, some pioneering insights into AD have also been offered by scholars adopting the quantitative approach. Experimental cognitive and psychological studies, based on measuring eye movements, heart rate or electrodermal activity, have explored the influence of AD on experience, presence and emotions (Fryer 2013; Walczak & Fryer 2017; Matamala et al. 2020), and accessibility of visual experience through AD (Holšánová 2008; 2016).

Most recent fundamental findings, to which this study refers, have also come from the realms of corpus and cognitive linguistics. Using transcribed AD corpora, scholars have confirmed the presence of subjective language (Soler Gallego 2019; Soler Gallego & Luque Colmenero 2018), extensive use of metaphor (Luque Olmenero & Soler Gallego 2020; Spinzi 2019; Bartolini 2023) and high lexical density (Perego 2019); discovering, amongst others, that museum AD discourse is lexically and grammatically richer than expected and more abundant in metaphor than academic, fictional or conversational discourses, displaying relatively high textual complexity (Perego 2024). They also confirmed that “the artistic style could have an impact on the audio description” and that descriptions related to works characterized by “higher levels of conceptuality and abstraction” encourage the use of subjective and figurative language (Luque Colmenero & Soler Gallego 2020).

Material and research questions

Our research builds upon these quantitative findings. While scholars to date have asked questions concerning the discourse-specific features of the language of AD (the art sources they refer to grouped at best into “representative”, “semi-abstract” and “abstract art” categories), we propose to explore how and to what extent the language of AD is motivated by the visual style it “translates” and how it can potentially be used to enrich or otherwise affect the sensory, aesthetic, emotional, and cognitive experience of a given artwork belonging to a different sensory modality.

Following de Coster and Mühleis's (2007) definition of museum AD as an act of "translating the visual sensation of works of art that museum visitors cannot touch", in which describers need to "explore the field of intersensorial possibilities" to help recipients form vivid mental representations of works of art (which are oftentimes visually ambiguous and complex), we analyze specific facets of this "intersensorial translation" by looking at embodied semantic attributes most salient in the descriptions of artworks belonging to specific art movements and individual artists' oeuvres.

To this end, we explore a bilingual corpus of 250 Anglophone and Polish pre-recorded and transcribed museum audio descriptions, made available on museum websites worldwide (these include such cultural institutions as MOMA, the MET, Tate Modern, National Gallery of Australia, The National Museum in Warsaw, Łódź and Wrocław) that refer to five selected nineteenth and early-twentieth century art movements: Impressionism, Expressionism, Abstractionism, Cubism and Surrealism.

In a quantitative analysis, we test:

1. To what extent do audio descriptions of paintings from various artistic movements differ in the utilized verbal imagery? Which features dominate in each movement/style (e.g., specific sensory modalities, emotional evaluation, etc.).
2. How do part-of-speech-specific patterns (adjectives, adverbs, nouns, verbs) in different semantic dimensions vary across movements?
3. Do these linguistic patterns reflect stylistic priorities of the visual art movements (e.g., color and light in Impressionism, motion and emotion in Expressionism, analysis/synthesis, and spatial disruption in Cubism)?
4. Are there systematic differences between Polish and English ADs in how they mediate visual art through language?

Methodology

a) Corpus and Data Preprocessing

Each of the 250 transcribed audio descriptions in the corpus (balanced across the five art movements and the two languages) is POS-tagged and lemmatized using the spaCY Python package. Relative frequencies of each group of content words (nouns, verbs, adjectives, and adverbs) are computed per text and normalized by its total word count.

b) Semantic Feature Rating via Transformer Models

To quantify the semantic content of each word, we leverage a brain-based componential semantic framework developed by Binder et al. (2016), proposing a neurobiologically grounded model of word-meaning composed of 65 experiential attributes derived from large-scale functional divisions in the human brain. These attributes span sensory, motor, spatial, temporal, affective, social, and cognitive domains.

The original dataset consists of human-generated ratings (0–6 scale) for 535 common English nouns, verbs, and adjectives across all 65 dimensions. These ratings reflect the degree to which each concept evokes a given experiential attribute, providing insights into how language activates perceptual and emotional systems.

To extend this framework to the full vocabulary of our AD corpus, we fine-tuned transformer models from the BERT family (Devlin et al., 2019) to predict Binder ratings for unseen words using sentences containing the original 535 seed words and their corresponding feature ratings in both Polish and English texts. For each content word in the AD corpus, the fine-tuned BERT model outputs a 65-dimensional semantic feature vector (ratings 0–5). The ratings are averaged for each part of speech (adjectives, adverbs, nouns, verbs) in each AD text. This makes it possible to assess specific patterns, capturing the relative experiential emphasis characteristic of each artistic movement.

c) Statistical Modeling: Generalized Linear Models (GLMs)

To identify the semantic dimensions and parts of speech (per-word frequency) most sensitive to artistic movement, generalized linear models are fitted separately to each of the 260 feature-by-POS combinations (65 Binder dimensions × 4 parts of speech) as well as the 4 POS frequency variables, with *Movement* entered as a categorical predictor. Model families and link functions are selected according to the distributional properties of each feature. The strength of the *Movement* effect is

assessed through a combination of statistical significance and model fit: (1) omnibus p-values from likelihood-ratio tests comparing the full model against a null (intercept-only) model are subjected to false discovery rate (FDR) correction, and (2) model improvement is quantified by the difference in Akaike Information Criterion (ΔAIC) between the same model pair. Features are ranked jointly by FDR-adjusted significance and ΔAIC magnitude. The highest-ranking dimensions undergo detailed modeling, including estimation of movement-specific effects, confidence intervals, and adjusted pairwise contrasts. This procedure isolates the most robust linguistic correlates of visual style and enables interpretation of their specific semantic content.

Expected Contribution

Our study aims to advance research in three interconnected domains:

a) Aesthetics and Embodied Cognition

The identifiable correspondences between artistic movements and embodied semantic profiles might suggest that aesthetic experience is not merely visual but cross-modally reconstructible. Linguistic mediation may thus preserve—and selectively amplify—perceptual and affective priorities embedded in visual style.

b) Accessibility and Professional Practice

By identifying which experiential dimensions are amplified or underrepresented in AD across movements, our findings may inform evidence-based refinement of AD guidelines. Rather than prescribing uniform descriptive strategies, accessibility practice could adopt style-sensitive approaches that better convey aesthetic intent.

c) Corpus-Based Art Studies

Methodologically, we introduce a scalable framework for modeling intersensorial translation using transformer-based semantic prediction combined with rigorous statistical modeling. The approach is extendable to additional museums, languages, and art movements, offering a replicable architecture for large-scale comparative research.

Ultimately, the study asks whether artistic style leaves a measurable semantic trace in language and what that implies for the reconstruction of aesthetic experience across sensory modalities.

References

Bartolini, C. 2023. “Museum Audio Descriptions vs. General Audio Guides: Describing or Interpreting Cultural Heritage?” Journal of Audiovisual Translation, 6, 77–98.

Binder, J. R., Conant, L. L., Humphries, C. J., Fernandino, L., Simons, S. B., Aguilar, M., & Desai, R. H. 2016. “Toward a Brain-Based Componential Semantic Representation.” Cognitive Neuropsychology, 33(3-4), 130–174.

De Coster, K. & Mühleis, V. 2007. “Intersensorial Translation. Visual Art Made up by Words” in Media for All: Subtitling for The Deaf, Audio Description and Sign Language (eds. Diaz Cintas, J., Orero, P. & Remael, A.) 189–200. Rodopi.

Devlin, J., Chang, M., Lee, K., & Toutanova, K. 2019. “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.” In Proceedings of the conference of the North American chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT), 4171–4186.

Fryer, L. 2013. Putting It into Words: The Impact of Visual Impairment on Perception, Experience and Presence. Unpublished doctoral thesis, Department of Psychology, Goldsmiths College, University of London. https://research.gold.ac.uk/id/eprint/10152/1/PSY_thesis_Fryer_2013.pdf

Holšánová, J. 2008. Discourse, Vision and Cognition. John Benjamins.

Holšánová, J. 2016. “A Cognitive Approach to Audio Description” in Researching Audio Description: New Approaches (eds. Matamala, A. & Orero, P.) 49–73. Palgrave.

Hirvonen, M. & Wiklund, M. 2021. “From image to text to speech: The effects of speech prosody on information sequencing in audio description”. Text and Talk 41, 309–334.

Honnibal, M. & Montani, I. 2017. “spaCy 2: natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing.”

Jerzakowska, B. 2016. Posłuchać obrazów. Podręcznik z audiodeskrypcją do reprodukcji malarskich, uzupełniający kształcenie literackie i językowe uczniów niewidomych. Rys.

Luque Olmenero, M. O. & Soler Gallego, S. 2020. “Metaphor as Creativity in Audio Descriptive Tours for Art Museums: From Description to Practice”. Journal of Audiovisual Translation 3, 64–78.

Matamala, A., Soler-Vilageliu, O. & Méndez-Ulrich, J.-L. 2020. “Electrodermal activity as a measure of emotions in media accessibility research: methodological considerations.” The Journal of Specialized Translation 33.

Neves, J. 2010. “Sound painting: audio description in another light.” Paper presented at the 8th International Conference & Exhibition on Language Transfer in the Audiovisual Media – Languages & The Media.

Neves, J. 2012. “Multi-Sensory Approaches to (Audio) Describing the Visual Arts” in MonTI. Monografías de Traducción e Interpretación, 277–293.

Pawłowska, A. & Sowińska-Heim, J. 2016. Audiodeskrypcja dzieł sztuki: metody, problemy, przykłady. Wydawnictwo Uniwersytetu Łódzkiego.

Pawłowska, A. & Sowińska-Heim, J. 2019. Osoby z niepełnosprawnością i sztuka. Udostępnianie, percepcja, integracja. Wydawnictwo Uniwersytetu Łódzkiego.

Perego, E. 2019. “Into the language of museum audio descriptions: a corpus-based study”. Perspectives: Studies in Translation Theory and Practice 27, 333–349.

Perego, E. 2024. Audio Description for the Arts: A Linguistic Perspective. Routledge.

Reviers, N. & Hanoulle, S. 2023. “Aesthetics and Participation in Accessible Art Experiences: Reflections on an Action Research Project of an Audio Guide”. Journal of Audiovisual Translation 6, 99–120.

Soler Gallego, S. 2019. “Defining Subjectivity in Visual Art Audio Description”. Meta 64(3), 708–733.

Soler Gallego, S. & Luque Colmenero, M. O. 2018. “Paintings to My Ears: A Method of Studying Subjectivity in Audio Description for Art Museums”. Linguistica Antverpiensia New Series 17, 140–156.

Soler, S., Kaleidoscope, Olalla, A. M. & Colmenero, L. 2023. “Increased Subjectivity in Audio Description of Visual Art: A Focus Group Reception Study of Content Minimalism and Interpretive Voicing”. Journal of Audiovisual Translation 6, 55–76.

Spinzi, C. 2019. “A cross-cultural study of figurative language in museum audio descriptions: implications for translation.” Lingue e Linguaggi 33, 303–316.

Vizcaíno, G. 2023. “Audio Description in Abstract Art: Using Metaphors From a Functional Perspective”. Journal of Audiovisual Translation 6, 189–208.

Walczak, A. & Fryer, L. 2017. “Creative Description: The Impact of Audio Description Style on Presence in Visually Impaired Audiences”. British Journal of Visual Impairment 35, 6–17.

Więckowski, R. 2014. “Audiodeskrypcja piękna”. Przekładaniec 28, 109–123.

14:30–14:50 *SHORT PAPER* **[7]**

Application of Multimodal AI in Classifying and Researching Zenta Dzividzinska's Photo Archive

Anda Baklāne[1], Valdis Saulespurēns[1], Alise Tifentale[2]

[1] *National Library of Latvia, Latvia*

[2] *City University of New York, Kingsborough Community College*

Keywords: *multimodal AI, computer vision, digital art history, photographic archives, metadata enrichment*

The paper investigates the potential of multimodal artificial intelligence for the exploration and interpretation of large-scale photographic collections, focusing on the development of a framework for the characterization and visualization of an author’s stylistic and artistic identity. It is based on the ideas of machine-learning-based computational image analysis (Wasllievski 2023; Chumachenko 2023; Guhennec 2024), cultural analytics (Manovich 2020), and distant viewing (Arnold 2019), touching upon the changes introduced by the use of large multimodal models in the analysis of images in art history (Impett 2024; Nygren 2023; Smits 2023). The project explores how computer vision-based methodologies, including automated image recognition, clustering, and multimodal text-image retrieval, can be applied to describe and compare recurring visual features across an individual photographer’s body of work. The approach aims to establish a foundation for identifying both topic-based and form-based regularities, thereby supporting a systematic, data-driven understanding of creative practice.

As a case study, we apply these methods to the largely underexplored archive of Zenta Dzividzinska (1944–2011), a Latvian artist and photographer whose work has begun to receive scholarly attention only posthumously. Produced primarily during the 1960s and 1970s in Soviet Latvia, Dzividzinskas's photographic negatives reveal a distinctive documentary sensibility centered on women’s everyday labour and domestic life – activities such as childrearing, food preparation, and agricultural work – that stood in marked contrast to the dominant, male-authored visual culture of the period. Her diverse representations of women as active agents rather than idealized objects constitute a subtle but

significant departure from the gendered conventions of Soviet photography. By integrating multimodal AI methods into the study of this corpus, the project seeks both to advance computational approaches to stylistic analysis and to enhance the visibility and scholarly recognition of Dzividzinska's work within Latvian photographic history through publications and AI-based tools for exploring the collection. This conference paper presents part of the ongoing research and emphasizes the study's technical aspects rather than its interpretive dimensions.

In the initial stage of the project, the authors explored the functionalities of the Collection Space Navigator (CSN), developed under the auspices of Tallinn University (Ohm 2023) (see https://zenta.lnb.lv/), before transitioning to a multimodal model-based application designed in-house at the National Library of Latvia (see https://zenta.lnb.lv/clip/multi). The CSN serves as an interface for exploring image collections via both metadata and visual embeddings generated by the ResNet-50 convolutional neural network. The metadata layer may include thematic keywords assigned either manually by researchers or automatically by AI models. While the similarity-based image cloud within the CSN environment provides a valuable means of visual browsing across the collection, it remains limited in its analytical utility without further methodological elaboration and interpretive structuring.

In the subsequent stage, the workflow was restructured around the CLIP (Contrastive Language-Image Pre-training) multimodal model (Radford 2021), with the goal of developing a framework for characterizing and visualizing an author's stylistic and artistic identity. To facilitate this, three categories of descriptive keywords were established: (1) thematic I, referring to the objects and subjects depicted within the frame; (2) thematic II, capturing the atmospheric and affective qualities of the image; and (3) formal, denoting visual and technical parameters such as lighting, camera angle, and compositional structure. The two thematic layers of keywords were generated by processing the collection with the Gemini language model, while the framework for formal characteristics was designed and refined by the authors of this paper.

In developing the framework, we tested and compared the descriptive categories and keyword sets originally assigned by the archivist, who created the collection's initial metadata, with several AI-generated keyword sets produced through automated processing.

As part of metadata enrichment, all images were embedded with XMP metadata containers that encompassed the Dublin Core and IPTC namespaces, as well as custom fields. Metadata enrichment posed additional challenges, ensuring that various tag and caption embeddings could be processed by the widest range of photographic tools.

In the final stage of the project, the oeuvre of Zenta Dzividzinska was represented through a series of visualizations that mapped the established thematic and formal categories. This stage was conceptually inspired by the Selfiecity project, which employed image plots to visualize photographic data along parameters such as city, gender, and mood (Tifentale 2015). In the case of Dzividzinska's archive, the image plots were generated based on a framework of thematic and formal characteristics developed through collaboration between researchers and AI tools, enabling a visual synthesis of the photographer's stylistic and artistic profile.

The project began as an exploration of CNN-based image recognition technologies and later evolved to the use of multimodal models such as CLIP. This methodological shift prompted reflection on the theoretical implications of moving from non-verbal models – those that classify and cluster images according to visual features that do not necessarily align with the interpretive frameworks developed by human researchers enables open-ended, language-based querying of collections and, to some extent, brings us enables open-ended, language-based querying of collections and, to some extent, brings toward text-aligned models. The latter enables open-ended, language-based querying of collections and, to some extent, returns the field from the discovery of previously unknown visual patterns to human-centered interpretive frameworks, reintroducing the subjectivity of language into the process of computational image analysis.

The provisional findings allow us to characterize Dzividzinska's collection: she uses more high-, low-, and Dutch-angle shots than eye-level shots; more medium- and wide-angle shots than close-ups; and her work is human-centered, with a particular focus on women.

Resources:

Collection Space Navigator instance for Zenta Dzividzinska's archive: https://zenta.lnb.lv/

Web application enabling access to the multimodal CLIP model: https://zenta.lnb.lv/clip/multi

Artist's webpage presenting the works of Zenta Dzividzinska: https://www.artdays.net/zdz

References

Arnold, T., & Tilton, L. (2019). Distant viewing: Analyzing large visual corpora. Digital Scholarship in the Humanities, 34(Suppl_1), i3–i16. https://doi.org/10.1093/llc/fqz013

Chumachenko, K., Männistö, A., Iosifidis, A., & Raitoharju, J. (2023). Machine learning–based analysis of Finnish World War II photographers. IEEE Access, 11, 124812–124826. https://doi.org/10.48550/arXiv.1904.09811

Guhennec, P., & Charlesworth, E. (2024). The computational eye: Deconstructing style in digital art history. Artl@s Bulletin, 13(2), Article 9. https://docs.lib.purdue.edu/artlas/vol13/iss2/9

Impett, L., & Offert, F. (2022). There is a digital art history. Visual Resources, 38(2), 186–209. https://doi.org/10.1080/01973762.2024.2362466

Manovich, L. (2020). Cultural analytics. MIT Press.

Nygren, C. (2023). Art history and AI: Ten axioms. Journal for Digital Art History, 9. https://doi.org/10.11588/DAH.2023.9.90400

Ohm, T., Canet Solà, M., Karjus, A., & Schich, M. (2023). Collection Space Navigator: An interactive visualization interface for multidimensional datasets. arXiv. https://doi.org/10.48550/arXiv.2305.06809

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., & Sutskever, I. (2021). Learning transferable visual models from natural language supervision. arXiv. https://arxiv.org/abs/2103.00020

Smits, T., & Wevers, M. (2023). A multimodal turn in digital humanities: Using contrastive machine learning models to explore, enrich, and analyze digital visual historical collections. Digital Scholarship in the Humanities, 38(3), 1267–1280. https://doi.org/10.1093/llc/fqad008

Tifentale, A., & Manovich, L. (2015). Selfiecity: Exploring photography and self-fashioning in social media. In D. M. Berry & M. Dieter (Eds.), Postdigital aesthetics: Art, computation and design (pp. 109–122). Palgrave Macmillan.

Wasielewski, A. (2023). Computational formalism: Art history and machine learning. MIT Press.

14:50–15:10 *LONG PAPER* **[8]**

Letters, Ledgers, and Lives: An Armenian Photographer's Archive and the Ethics of Digital Curation

Idil Cetin

University of Oslo, Norway

Keywords: *Armenian diaspora, digital archives, ethics of curation, archival silences*

This paper uses the personal archive of Garabet Chahinian, an Armenian photographer who was active in early 20th-century Istanbul, as a case study in curating non-canonical heritage within digital frameworks. The collection consists of letters sent to Chahinian from Armenians across the diaspora, his responses which he copied into several notebooks, and the detailed account books he kept for his studio. The collection forms a central part of my wider project at the University of Oslo, "Photography as a Survival and Diaspora-Building Tool in the Aftermath of 1915."

My encounter with Chahinian's collection began almost five years ago, when a handful of his letters appeared on Turkish auction sites. Believing that they had ended up in the wrong hands, I purchased them with the intention of returning them to the family and finding out more about Chahinian, who, I discovered, became a photographer in his late forties. Over time, hundreds of additional documents appeared for sale, and I soon discovered that the family itself was selling the collection. With the support of another branch of the family, who gave me their blessing, I incorporated this material into my research, where it now serves as a lens onto the Armenian diaspora experience after the genocide.

The letters trace the dispersal of Chahinian's extended network across multiple geographies, recording resilience and survival strategies as well as ongoing ties to "home." Meanwhile, his account books reveal that the studio primarily served Istanbul's shrinking non-Muslim communities in the early Republican period. Written largely in Western Armenian, which is classified as an endangered language, the collection is being translated into English and encoded in XML-TEI in order to develop a searchable digital database. In doing so, the project not only makes an overlooked survivor's story

accessible to a global audience but also demonstrates how scattered diaspora archives can counterbalance the gaps and biases of institutional collections.

In this paper, my focus will be on the ethical challenges of bringing such a sensitive archive into public view. Chahinian's letters chronicle deeply personal and communal struggles in the aftermath of 1915 and require careful handling. How can we avoid sensationalising trauma or reducing complex lives to narratives of suffering? How can we preserve the intimate nature of correspondence when integrating it into digital libraries? Most importantly, how can we respect the consent and privacy of families and communities, particularly given that the collection represents not a single individual, but an interconnected web of voices?

Foregrounding responsible storytelling also means recognising the ethical responsibility to make such stories known. Armenian history is characterised by ruptures, dispersals, and archival absences. To allow Chahinian's documents to disappear into obscurity would mean perpetuating those silences. Therefore, the act of research and digital preservation is not only a matter of scholarly rigour but also an ethical obligation to ensure ensure that vulnerable histories are safeguarded and shared, rather than lost.

Combining cultural sensitivity with digital tools, this paper proposes best practices for curating non-canonical archives, such as engaging community members in shaping metadata and annotations, framing contextual narratives that foreground the subject's agency, and developing digital platforms that educate and engage without resorting to voyeurism or over-simplification. The ultimate aim is to demonstrate how the ethical curation of "lost" personal archives can empower marginalised histories to be told on their own terms, and how making them visible constitutes both an academic contribution and a moral responsibility.

Session 1C — 13:30–15:30

13:30–13:50 **The perception of song within the hymn material of N. F. S. Grundtvig**
Anna Pouline Brogaard

13:50–14:20 **Nordic Places of Worship (NordPoW) – an example of how to use GIS-map to document and visualize religious geographies and the neglected cultural heritage of prayer houses**
Stefan Gelfgren, Jakob Dahlbacka, Bo Ejstrud, Andreas Tjomsland

14:20–14:40 **Preaching During the COVID-19 Pandemic: A Quantitative Close Reading of Danish Sermons during National Lockdowns**
Emil Walther Bønding, Michael Mørch Thunbo, Anne Agersnap

14:40–15:10 **From Flat Data to Deep History: A New Testament Corpus for the Comparative Humanities**
Maciej Rapacz

15:10–15:30 **Modelling the Mythological Structure "Birth–Life–Death–Immortality" in Ukrainian and Estonian Folk Songs through Zero-Shot and Embedding-Based Comparative Analysis**
Olha Petrovych, Mari Väina

13:30–13:50 *SHORT PAPER* **[9]**

The perception of song within the hymn material of N. F. S. Grundtvig

Anna Pouline Brogaard
Aarhus University, Denmark

Keywords: *Song, Hymns, Grundtvig, Annotation*

Every Sunday, Danish churches are filled with the sound of songs spanning much of Christian cultural history. While we are accustomed to music charts being dominated by the latest popular music, the Danish hymn-singing tradition emphasizes the resonance of history (Solten, 2014, p. 226). The hymn offers insight into changing Christian worldviews and approaches to existential questions, while at the same time functioning as a practical text, ritually grounded in and indispensable to the liturgy of the Church of Denmark (Solten, 2014, p. 227). The relevance and significance of the hymn tradition are

particularly evident in the work of the national poet N. F. S. Grundtvig. According to Grundtvig, song was the central mode of Christian expression, and hymn writing was also the part of his authorship that gave his understanding of Christianity its greatest popular impact (Isaksen & Bak, 2018, p. 98; Solten, 2014, pp. 223-224).

Consequently, the purpose of this study is to map the semantic landscape of song terms in N. F. S. Grundtvig's hymns, as well as to revisit already established research and systematically test its theoretical frameworks. The theories concern the worldview embedded in Grundtvig's hymns, with particular attention to his world-affirming tendencies as well as the role of angels as mediators connecting heaven and earth (Baunvig, 2019; Baunvig & Nielbo, 2022). They also include the practical dimension of the hymn genre and Grundtvig's awareness of the hymn's capacity to foster communal identity and facilitate the internalization of shared beliefs (Baunvig, 2013). My project seeks to contribute a systematic approach grounded in the study of religion by computationally analysing occurrences of song in Grundtvig's hymn corpus using the method of *Quantitative Close Reading* (Agersnap et al., 2025). A method that combines classical close reading practices and computational approaches through the use of formalized interpretation. The method highlights human intervention in automated processes, also referred to as Human-in-the-Loop (HITL) (Ibid., 2). My material consists of 531 hymns, including *Sangværk til den Danske Kirke* and a substantial sample of Grundtvig's hymns from the Danish Hymnal.

To explore the role of song within this material, I developed an annotation manual describing the annotation process, focusing on song terms as the selected keyword. First, occurrences of the keyword are identified, after which the surrounding context is annotated according to the categories defined in the annotation manual. These categories are developed based on the theoretical framework mentioned above. In this way, the manual facilitates a systematic close reading of the hymns, after which the categories are translated into a quantitative parameter in the form of binary coding (Agersnap et al., 2025, pp. 4–5). Taken together, this constitutes a quantitative close reading.

The annotation process involves a considerable degree of interpretation, particularly when assessing more abstract categories in poetic material. For this reason, I conducted an annotation test to measure inter-observer reliability (Landis & Koch, 1977). The test showed a moderate to substantial level of annotator agreement, ensuring the replicability, generalizability, and validity of the study's results. In addition to manual annotation, I included model-assisted annotation, where I trained a ChatGPT agent to annotate according to the study's objectives and the coding manual's guidelines and to cross-reference it with already annotated hymns. The model-based annotation was subsequently quality-assured through systematic sampling, during which the model's choices were verified and adjusted. Human formalization is considered the most effective way to secure consistent and reliable model-assisted text annotation (Kristensen-McLachlan et al., 2025, p. 3) The overall dataset thus rests on a combination of human interpretation and model assistance.

Since the method enables not only systematic pattern recognition but also scalable analysis, my interpretation of the results allows for more extensive argumentation incorporating complex theories from the study of religion. For example, the associations and embedding of the song terms illustrate the praise of God's joyous creation of the earth relates to Grundtvig's world-affirming worldview. They also emphasise the collective practise of hymn singing. The analysis confirms established theoretical frameworks and demonstrates the semantic function of song in Grundtvig's hymns. The song terms connect practical use with narrative content, thereby linking the world of the singer with the world of what is sung. In this way, the hymn unites a concrete moment in time with a cosmological, creating an ontic resonance between heaven and earth.

References

Agersnap, A., Schmidt, L. W., Eriksen, R. S., Bønding, E. W., Kirkegaard, T. H., Borčak, L. W., & Baunvig, K. L.

2025. The Human Touch: Leveraging HITL for Quantitative Close Reading of Historical Corpora. Digital Humanities in the Nordic and Baltic Countries Conference (DHNB 2025), 9th.

Baunvig, K. F.

2013. Forsamlingen først : N.F.S. Grundtvigs og Émile Durkheims syn på fællesskab Aarhus Universitet]. Århus.

2019. Maddikerne ud af lyrikken: Af-tantrificering og genfortryllelse i N.F.S. Grundtvigs salmer. Religionsvidenskabeligt tidsskrift(69), 164-186. https://doi.org/10.7146/rt.v0i69.114475

Baunvig, K. F., & Nielbo, K. L.

2022. Mermaids are Birds : Embedding N.F.S. Grundtvig's Bestiary. In: CEUR-WS.org.

Isaksen, S., & Bak, K. S.
2018. Fællessang og fællesskab : en antologi (1. udgave. ed.). Videnscenter for Sang, Sangens Hus.
Kristensen-McLachlan, R. D., Canavan, M., Kárdos, M., Jacobsen, M., & Aarøe, L.
2025. Are chatbots reliable text annotators? Sometimes. PNAS nexus, 4(4), pgaf069. https://doi.org/10.1093/pnasnexus/pgaf069
Landis, J. R., & Koch, G. G.
1977. The Measurement of Observer Agreement for Categorical Data. Biometrics, 33(1), 159-174. https://doi.org/10.2307/2529310
Solten, T. B.
2014. Salmen. In C. S. Sune Auken (Ed.), Ved lejlighed. Grundtvig og genrerne (pp. 223-251). Spring.

13:50–14:20 *LONG PAPER* **[10]**

Nordic Places of Worship (NordPoW) – an example of how to use GIS-map to document and visualize religious geographies and the neglected cultural heritage of prayer houses

Stefan Gelfgren[1], Jakob Dahlbacka[2], Bo Ejstrud[3], Andreas Tjomsland[4]
[1] *Umeå University, Sweden*
[2] *Åbo Akademi University, Finland*
[3] *Holstebro Museum, Denmark*
[4] *Volda University College and University of Agder, Norway*

Keywords: *data visualization, GIS, religious geography, church history, data base buildning*

Using the so-called NordPoW (Nordic Places of Worship) map – a collaborative Nordic mapping of churches and prayer houses – as the starting point, (for a deeper understanding of the project see Tjomsland et al 2024) this paper explores the importance of dual competencies in Digital Humanities projects aimed at engaging traditional Humanist scholars and stakeholders. It argues that both the Humanist and the digital developer bring equally vital expertise to ensure project success. This might be obvious, but it is tied to a discussion on possibilities and limitations of both humanistic and computational methods and competencies. (compare Henriot 2023; Oberbichler et al 2022)

Ultimately, the research questions and data curation processes are grounded in our disciplinary backgrounds in the traditional humanities. The NordPoW project group consists of four researchers with expertise in sociology of religion, history, archaeology, theology, church history, and digital humanities. We have collaborated closely with system developers specializing in GIS from Humlab at Umeå University and the Swedish research infrastructure InfraVis. (for InfraVis see Weinkauf et al 2025) We argue that this project would not have been possible without strong competencies in both traditional humanities and computational methods. This combination enabled the development of a robust research infrastructure.

However, a well-defined research question is fundamental to any research-oriented Digital Humanities initiative. Regardless of technical proficiency – whether in system development, UX design, or software engineering – without a guiding research question, the project risks becoming irrelevant.

This argument is examined through experiences from building a database of churches and prayer houses, visualized on a GIS-based map. (cf. Stump 2008) Each building is represented with a timeline and filter function. The project began with mapping churches and prayer houses in northern Sweden (originally the so-called DigiBön map) and later expanded into a Nordic initiative covering Sweden, Denmark, Finland, and Norway—now known as NordPoW.

The Swedish dataset was compiled from inventories provided by museums, local authorities, the Church of Sweden, and the Swedish National Heritage Board. It was further enriched through web searches, public contributions, and Google Street View browsing. The data arrived in various formats, including Excel sheets, emails, scanned handwritten PDFs, and Word documents. Similar data collection processes were used in the other participating countries.

Due to processes often referred to as secularization (a concept widely debated but beyond the scope of this paper – see for example Berger 2014; Bruce 2011) the landscape of churches and prayer houses

is changing. While national churches such as the Church of Sweden and the Church of Norway are often legally protected and considered cultural heritage, prayer houses typically lack such protection. The NordPoW map thus also serves as a platform for preserving this diminishing cultural heritage.

Currently, the database contains approximately 6,000 entries, including all national churches in Sweden and Norway, all prayer houses in the Diocese of Luleå (the northern third part of Sweden), and all prayer houses in the Norwegian counties of Sogn og Fjordane and Møre og Romsdal. Although only a fraction of prayer houses are currently included at the national level, these regions are comprehensively covered. The total number of expected entries is estimated to be between 20,000 and 30,000. More regional data from Sweden and Norway are coming, and data from Finland and Denmark are on the way.

The database is visualized on an ArcGIS-based map, with each object represented by a point/symbol. It is designed to be scalable, allowing for the inclusion of additional regions and integration with other datasets, such as non-Chrisitan buildings, demographic data, voting patterns, or crime statistics.

The infrastructure is intended to support research primarily in religious geography across the Nordic countries. Consequently, the research questions are shaped by shared scholarly interests and disciplinary expertise. The project includes a wide range of churches and denominations, with both similarities and differences across national borders. Some movements – such as Pentecostalism and Laestadianism – exist under the same name across countries, while others, like the inner mission movement or Lutheran free churches, vary in terminology despite conceptual similarities. To enable cross-country comparisons, data must be merged and standardized, which requires specialized knowledge in the field of study.

Defining the type of research the infrastructure aims to support is therefore essential. In our case, we investigate questions related to religious geography – a classic field for historians, church historians, and sociologists of religion. (Tuan 1976; Bodenhamer 2015; Foka et al. 2020; Zhao 2022) The NordPoW project centres around questions such as: What does the religious geography of the Nordic countries look like over time? How do different religious affiliations relate to one another? Are there regions of strong or weak religiosity? How do these patterns evolve?

The map also serves as a point of entry for studying specific areas or phenomena related to these overarching questions. To achieve this, data must, as mentioned, be curated accordingly. In the NordPoW case, key parameters include religious affiliation, geographical coordinates, and the date associated with each building. (following Gustafsson 1957) While these may seem straightforward, each parameter involves negotiation and interpretation.

Religious affiliation is generally manageable, although historical collaborations and transformations can complicate categorization. The most challenging aspect is aligning national affiliations across borders. Location data may be missing, ambiguous, or complicated by buildings being relocated or congregations moving. Nevertheless, a definitive location must be established. Regarding dates, discrepancies often arise between the year a building was constructed and the year it was inaugurated by a denomination. These issues required extensive negotiation and input from external experts. (compare Nygren et al 2016).

In general, we have prioritized identifying patterns over achieving perfect accuracy. For the purposes of the map, minor deviations in location or date are acceptable. From a research infrastructure perspective, this level of precision is sufficient. However, for individuals personally connected to these churches and prayer houses – those who built them or grew up in them – accuracy is far more important. This presents a potential conflict of interest, which must be acknowledged and addressed. Public engagement is therefore crucial for quality control, and we revise the data based on incoming comments and suggestions. Hereby, the NordPoW map brings previously neglected data together, and visualize them in a way never done before.

A visually coherent and intuitive symbology must also be chosen to ensure that the patterns intended for visualization are not obscured by inconsistent colors or symbols. In other words, behind the seemingly coherent and binary digital map lie numerous negotiations and considerations—rooted in diverse and interrelated forms of expertise.

In conclusion, we emphasize the importance of linking research questions with appropriate methods. A Digital Humanities project should be equally grounded in computational and humanities expertise. (cf. Münster 2020) Research questions are formulated by scholars, and computational methods are

developed collaboratively in relation to the intended outcomes. Interaction and communication between researchers and system developers/GIS experts—on equal footing—are therefore of utmost importance.

References

Berger, P. (2014). The Many Altars of Modernity: Toward a Paradigm for Religion in a Plural-ist Age. Boston: Walter De Gruyter.

Bodenhamer, David J. (2015). “Narrating Space and Place.” In Deep Maps and Spatial Narratives, edited by David J. Bodenhamer, John Corrigan, and Trevor M. Harris, 7–27. Bloomington: Indiana University Press.

Bruce, S. (2011). Secularization: In Defence of an Unfashionable Theory. Oxford: Oxford University Press.

Foka, A., Cocq, C., Buckland, P., & Gelfgren, S. (2020). “Geovisualization as Method.” In Routledge International Handbook of Research Methods in Digital Humanities.

Gustafsson, B. (1957). Svensk kyrkogeografi: Med samfundsbeskrivning [Swedish Church Geography: With Denominational Descriptions]. Lund: Gleerup.

Henriot, C., & Armand, C. (2023). Beyond Digital Humanities Thinking Computationally: Position Paper.

Münster, S., & Terras, M. (2020). The visual side of digital humanities: a survey on topics, researchers, and epistemic cultures. Digital Scholarship in the Humanities, 35(2), 366-389.

Nygren, T., Frank, Z., Bauch, N. & Stener, E. (2016) “Connecting with the past: Opportunities and Challenges in Digital History”. In Research Methods for Creating and Curating Data in the Digital Humanities. Matt Hayler & Gabriele Griffin. Edinburgh: Edinburgh University Press.

Oberbichler, S., Boroş, E., Doucet, A., Marjanen, J., Pfanzelter, E., Rautiainen, J., ... & Tolonen, M. (2022). Integrated interdisciplinary workflows for research on historical newspapers: Perspectives from humanities scholars, computer scientists, and librarians. Journal of the Association for Information Science and Technology, 73(2), 225-239.

Tjomsland, A., Gelfgren, S., Ejstrud, B., & Dahlbacka, J. (2024). Revealing the Revivals: Towards a Common Nordic Framework for Digital Mapping of Places of Worship. Nordic Journal of Religion and Society, 37(2), 106–122.

Tuan, Y.-F. (1976). “Humanistic Geography.” Annals of the Association of American Geographers 66 (2): 266–276.

Weinkauf, T., Romero, M., Besançon, L., Ahlstedt, J., Berendt, F., Billger, M., … Ynnerman, A. (2025). InfraVis : the Swedish research infrastructure for visualization support. VisGap : The Gap between Visualization Research and Visualization Software. Presented at the 27th EG Conference on Visualization (VisGap2025), Luxembourg, June 2-6, 2025.

Zhao, B. (2022). “Humanistic GIS: Toward a Research Agenda.” Annals of the American Association of Geographers 112 (6): 1576–1592.

14:20–14:40 *SHORT PAPER* **[11]**

Preaching During the COVID-19 Pandemic: A Quantitative Close Reading of Danish Sermons during National Lockdowns

Emil Walther Bønding, Michael Mørch Thunbo, Anne Agersnap

Aarhus University, Denmark

Keywords: *Quantitative Close Reading (QCR), sermons, COVID-19 lockdowns, thematic analysis*

Sermon manuscripts serve as important documentation of an ongoing religious practice. In Denmark, sermons form a central part of the Sunday service, where pastors interpret prescribed biblical texts while addressing the current concerns of their congregations. In times of crisis, sermons become especially revealing: they expose how pastors negotiate their dual roles as theological authorities and interpreters of social experience. This study examines sermons delivered in Denmark during the COVID-19 pandemic (2020–2021). The analysis treats pastors’ weekly sermons as expressions of the Evangelical Lutheran Church of Denmark’s (ELCD) collective voice (Agersnap 2021) and explores how their thematic focus shifts throughout the course of the crisis.

The study forms part of an ongoing digitisation project initiated and managed by the cultural institution *Danske Taler* (Danish Speeches). Over a three-year period, *Danske Taler* is uncovering Danish sermons from the sixteenth century to the present day, digitising them and making them publicly available on its platform. As part of this effort, the institution provides data for a research project

investigating preaching during national crises in Denmark (Danske Taler 2025). The project targets three historical periods: the Schleswig Wars (1849–1851; 1864), the German Occupation (1940–1945), and the COVID-19 pandemic. The complete crisis corpus comprises 300 sermons written by Danish pastors.

We apply a *Quantitative Close Reading* (QCR) approach (Agersnap et al. 2025) to annotate thematic structures across the sermons. QCR is a human-in-the-loop method that relies on digital data generation without using automated computational analyses. Instead, annotators conduct a structured reading of the corpus by manually tagging themes and contextual elements. The approach is particularly useful when the material is poorly suited for automated text analysis—for instance, due to faulty Optical Character Recognition—or when working with smaller digital collections, such as the sermon corpus.

An annotation codebook is developed from an initial test sample of thirty sermons, read and annotated through an inductive, bottom-up process to identify recurring semantic, thematic, and contextual features. Seven main thematic categories emerge: theological themes (e.g., *faith, sin, salvation*); cosmology (e.g., *heaven, hell*); biblical motifs; collectives (e.g., *church, nation, family*); emotive responses (e.g., *fear, courage, loss*); and social functions (e.g., *consolation, calls for resistance*, or *appeals to social cohesion*). Each sermon is also annotated for its semantic positioning of both the crisis and the prescribed biblical text (c*entral, peripheral*, or *omitted*). Through iterative rounds of discussion and refinement, these emergent codes are formalised into a structured annotation manual comprising 75 distinct codes. This process ensures both empirical sensitivity and conceptual coherence, establishing a reliable framework for systematic annotation and producing a structured dataset that supports diachronic analysis without losing the interpretative nuance of close reading.

The COVID-19 corpus consists of more than 100 sermons written by 51 pastors in the ELCD. The material is divided into two groups: sermons delivered during national lockdowns and sermons delivered during periods of reopening. This comparison highlights differences between times when the crisis directly impacted daily life and restricted freedom of action against times when the consequences of the crises affected public life to a lesser extent. The thematic analysis of the sermons shows significant shifts between the two societal conditions. During lockdowns, themes of care (*consolation, love, hope*) and the emotional and direct impact of the crises (*fear, loss, death,* and *the state of society*) are strongly emphasized. In contrast, abstract discussions of humanity—such as *existence, humility,* and *selfishness*—fade into the background during lockdowns but are heightened during reopenings. Despite these shifts, the larger cosmological framework remains stable: core ideas about *heaven, evil forces, the church,* and *light and darkness*, as well as theological themes, persist throughout. The findings suggest that, in their collective response to the Covid-19 crisis, the textual content of the Danish pastors' sermons moves between tending to the listener's immediate world during lockdowns and broad existential reflection during reopenings. The thematic shifts in the sermons repeat consistently during the recurring societal changes, while the texts remain firmly rooted in the collective narratives of the church throughout.

References

Agersnap, A. 2021. Collective Testimonies to Christianity and Time: A Collection and Large-Scale Text Study of 11,955 Danish Sermons from 2011-2016 [Doctoral disseration, Aarhus University].

Agersnap, A., L. Schmidt, R. Eriksen, E. Bønding, T. Kirkegaard, Lea Borcak & K. Baunvig. "The Human Touch: Leveraging HITL for Quantitative Close Reading". Proceedings of the 9th Digital Humanities in the Nordic and Baltic Countries Conference (DHNB 2025) (In press).

Danske Taler. 2025. "Nyt forskningsprojekt skal kaste lys over prædikener i krisetider". (January 20, 2025). https://www.dansketaler.dk/praedikener/nyheder/nyt-forskningsprojekt-skal-kaste-lys-over-praedikener-i-krisetider (seen October 29, 2025).

14:40–15:10 *LONG PAPER* **[12]**

From Flat Data to Deep History: A New Testament Corpus for the Comparative Humanities

Maciej Rapacz

AGH University of Kraków, Poland

Keywords: *Digital Humanities, Textual Scholarship, Natural Language Processing, Diachronic Analysis, Parallel Corpus*

1. Introduction

In early Natural Language Processing (NLP), the Bible was a widely used parallel corpus for tasks like statistical machine translation, mostly for two reasons: first, the fact that it is the most translated text and hence the most available, and second, it is already pre-structured into verses, which made it easier to form parallel corpora.

Today, as contemporary NLP models grow in terms of size and complexity, so do the training datasets needed to train them. At this scale, the utility of biblical corpora suggests a paradigm shift: from being a primary source of training data for AI systems to being an object of study *by* AI systems.

In the past, there have been numerous attempts at introducing digital biblical corpora, mostly designed for the first paradigm and, in turn, usually prioritizing *linguistic breadth* (a few translations across many languages) over *vertical depth* (many translations within one language). Their primary application — machine translation — relied on the assumption of verse-level semantic equivalence (Mueller et al., 2020); hence, multiple translations within one language were not needed and were usually treated as duplicates.

This "flat" approach, however, hides the complex revision history and stylistic variation within a single language, which is of primary interest to fields in digital humanities such as translation studies or linguistics. The notion of a "duplicate" is contingent on the research question. For a scholar studying the history of a single translation, such as the King James Version, each revision constitutes a unique work. For another aiming to compare translation strategies, multiple revisions of the same work might be considered duplicates.

Within our work, we introduce a novel corpus built with the second paradigm in mind. *targum*, a multilingual corpus of New Testament translations focused on vertical depth, comprising *657 translations* (352 unique works) spanning five languages: English, French, Italian, Polish, and Spanish.

We further manually tag each work with a standardized identifier (e.g. king-james-version), translation's edition (e.g. kjv-1611), and its year of revision (e.g. 1611). Instead of using heuristics to remove duplicates and imposing a single definition of "uniqueness" on the scholar, our metadata allows researchers to define it on their own.

2. Related Work

A significant body of work has focused on creating biblical corpora that maximize linguistic breadth. This trend includes foundational resources from Christodouloupoulos (2015) and Mayer (2014), and more recently the eBible Corpus (Akerman, 2023), culminating in collections that cover over 1,600 languages to support tasks such as typological analysis and low-resource NLP (McCarthy et al., 2020; Asgari and Schütze, 2017). While these collections offer unparalleled breadth, their depth for any single language remains limited.

In contrast, a few projects have curated collections with depth for specific scholarly questions. A prime example is the EDGeS Diachronic Bible Corpus, a resource containing 36 translations across four Germanic languages designed to study complex verb constructions (Bouma et al., 2020). Similarly, Carlson (2018) compiled 34 English versions for the specific task of evaluating the transfer of prose style. Our work seeks to *extend and generalize this depth-focused approach*. Although these valuable prior corpora were built to answer a predefined research question, targum is designed as a *general-purpose resource for comparative analysis*. Its aim is to allow researchers to dynamically create a wide variety of bespoke subcorpora tailored to their own specific inquiries.

3. Methodology

3.1. Data Acquisition

We aggregated translations from a diverse set of sources, including large-scale aggregators (e.g., biblegateway.com) and language-specific repositories (e.g., bibliepolskie.pl). The acquisition pipeline involved three stages:

1. *Indexing*, which involved inventorying all available translations on each source website.
2. *Scraping*, where we downloaded the full text of each New Testament, caching the raw source files (e.g., HTML, JSON) to ensure reproducibility.

3. *Parsing*, which processed the cached files to extract verse-level text, handling structural variations such as non-standard segments (e.g., 6a, 6b) or verse ranges (e.g., 6–8).

At the end of this stage, we had a total of 657 translations, each parsed into a structured format with verse-level text.

3.2. Metadata Annotation

The scraped data was inconsistent across sources, with the same translation often appearing under different names or with incorrect publication years. To resolve this, we performed a fully manual canonicalization process.

We initially explored using AI assistants for automated metadata extraction, but this approach proved unreliable, frequently failing to distinguish between closely-related revisions. Consequently, the final canonicalization was performed entirely by hand. Each of the 657 scraped instances was manually researched and mapped to a standardized entry stored in YAML files. Each entry contains:

- A canonical_id to identify a distinct translation work (e.g. king-james-version).
- A canonical_version to specify the exact edition (e.g. kjv-1611).
- A canonical_revision_year to specify the exact year of revision (e.g. 1611).
- A canonical_version_reasoning documenting the rationale for the classification. (e.g. "lexical match with tagged revision in the corpus")

This annotation process required a multi-faceted verification strategy, as source metadata was often incomplete. In straightforward cases, we could rely on the metadata provided directly by the source website. In others, we relied on contextual knowledge, such as the fact that a particular translation only ever had a single published edition. For more ambiguous cases, we inferred publication details from copyright statements or, in the most challenging instances, performed direct textual comparisons against scanned historical documents. Once a version was confidently identified, we propagated its metadata to any exact lexical duplicates found across other sources.

4. Corpus Analysis

The analysis of the corpus demonstrates its potential for new quantitative research by revealing patterns in its scale, historical distribution, and internal textual structure.

4.1. Corpus Scale and Comparative Depth

The primary contribution of targum is a vertical depth that substantially exceeds previous resources. We have gathered:

- English: 396 total, 208 unique translations
- French: 78 total, 41 unique
- Italian: 33 total, 18 unique
- Polish: 48 total, 30 unique
- Spanish: 102 total, 55 unique

Compared to the next largest resource (Mayer, 2014), this represents a significant increase in depth: 5.3x for English, 5.0x for Polish, 2.5x for Spanish, 2.6x for Italian, and 2.4x for French.

4.2. Diachronic Distribution

This numerical depth translates into a wide historical range, with translations spanning from the 16th-century Reformation period to the present day. The diachronic data reveals biases and gaps in the digital record: while the English collection shows a rather continuous history, the analysis uncovers a "digital dark age" for other languages: a near-complete absence of digitized translations for over 200 years between the early 17th and late 19th centuries in the Polish and Spanish collections. This 'digital dark age' reflects a digitization bias, as historical publications from this era exist but are not yet part of the online record. In all cases, the corpus is heavily skewed towards the modern period (mean: 2000). This gap limits cross-linguistic comparability for pre-modern periods and introduces discontinuities in diachronic analyses, particularly for Polish and Spanish. While part of this skew reflects a genuine historical trend, researchers should treat the corpus's historical coverage as asymmetric across languages.

4.3. Similarity Analysis

Finally, we computed pairwise similarity scores between all translations within each language using both a semantic approach (cosine similarity on Qwen3-Embedding-0.6B embeddings, selected for its state-of-the-art multilingual performance and long-context capabilities; Zhang et al., 2025) and a lexical approach (Levenshtein distance). The analysis revealed that while most versions are semantically close — with similarity scores being tightly clustered (mean: 0.88, std: 0.10) — the lexical similarity scores show greater variance and a lower mean (mean: 0.70, std: 0.12), proving more effective at distinguishing between closely related versions. This highlights a key methodological insight: a reliance on semantic similarity alone can obscure the fine-grained textual variations that are often the primary object of study in the digital humanities.

5. Discussion

The targum corpus and its metadata structure enable research questions that were previously difficult to address at scale.

Micro-Level Analysis: Researchers can select focused subsets of the corpus to perform targeted comparisons. For example, one can:

- Conduct diachronic studies by tracing linguistic changes across decades of revisions within a single translation family (e.g., the King James Version).
- Perform synchronic comparisons of contemporary versions to isolate the effects of confessional tradition or translator style.
- Track translation lineage by following the propagation of stylistic choices through indirect translation chains (e.g., those mediated by a Latin Vulgate source or a modern English pivot translation).

Macro-Level Analysis: A new translation, whether human- or machine-produced, can be embedded and situated within the semantic space of the entire corpus, allowing its stylistic and theological profile to be quantitatively compared against hundreds of historical precedents.

6. Conclusion

We introduce *targum*, a multilingual New Testament corpus aiming to maximize representation of translation history of the New Testament within the five selected European languages. By gathering 657 translations and, most importantly, manually annotating all of them into a standardized format, we provide a general-purpose resource for quantitative research on the interplay of theology, style, and culture in translation history.

The corpus and all associated metadata will be made publicly available at https://github.com/mrapacz/targum-corpus.

References

Akerman, V., Baines, D., Daspit, D., Hermjakob, U., Jang, T., Leong, C., Martin, M., Mathew, J., Robie, J., and Schwarting, M. (2023). The ebible corpus: Data and model benchmarks for bible translation for low-resource languages. *arXiv preprint arXiv:2304.09919*.

Alexander, P. S. (1992). Targum, Targumim. In *The Anchor Bible Dictionary*, 6, pp. 320-331.

Asgari, E. and Schütze, H. (2017). Past, Present, Future: A Computational Investigation of the Typology of Tense in 1000 Languages. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pp. 113-124.

Bouma, G., Coussé, E., Dijkstra, T., and van der Sijs, N. (2020). The EDGeS diachronic bible corpus. In *Proceedings of the 12th Language Resources and Evaluation Conference*.

Carlson, K., Riddell, A., and Rockmore, D. (2018). Evaluating prose style transfer with the bible. *Royal Society open science*, 5(10), p.171920.

Christodouloupoulos, C. and Steedman, M. (2015). A massively parallel corpus: the bible in 100 languages. *Language resources and evaluation*, 49(2), pp.375-395.

Mayer, T. and Cysouw, M. (2014). Creating a massively parallel Bible corpus. *Oceania*, 135(273), p.40.

McCarthy, A. D., Wicks, R., Lewis, D., Mueller, A., Wu, W., Adams, O., Nicolai, G., Post, M., and Yarowsky, D. (2020). The Johns Hopkins University Bible corpus: 1600+ tongues for typological exploration. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pp. 2884-2892.

Mueller, A., Nicolai, G., McCarthy, A. D., Lewis, D., Wu, W., and Yarowsky, D. (2020). An Analysis of Massively Multilingual Neural Machine Translation for Low-Resource Languages. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pp. 3710-3718.

Schuhlein, F. (1912). Targum. In *The Catholic Encyclopedia*, 14.

Zhang, Y., Li, M., Long, D., Zhang, X., Lin, H., Yang, B., Xie, P., Yang, A., Liu, D., Lin, J., Huang, F., and Zhou, J. (2025). Qwen3 Embedding: Advancing Text Embedding and Reranking Through Foundation Models. *arXiv preprint arXiv:2506.05176*.

15:10–15:30 *LONG PAPER* **[13]**

Modelling the Mythological Structure “Birth–Life–Death–Immortality” in Ukrainian and Estonian Folk Songs through Zero-Shot and Embedding-Based Comparative Analysis

Olha Petrovych, Mari Väina

Estonian Literary Museum, Estonia

Keywords: *semantic modelling, mythological structures, zero-shot classification, multilingual embeddings, oral poetry*

This study investigates how the mythological structure “birth → life → death → immortality” is linguistically and semantically represented in Ukrainian and Estonian folk songs, and how transformer-based computational semantics can reveal cross-cultural continuities and divergences in this archetypal framework. Drawing from psychoanalytic theories of vitality and mortality (Freud 1920; Heimann 1952), thanatological and anthropological perspectives on death as transformation (Feifel 1959; Ariès 1977), and structuralist and symbolic interpretations of myth (Lévi-Strauss 1966; Eliade 1957; Campbell 1949), this research explores how Ukrainian and Estonian folk songs encode the existential cycle of human experience within culturally specific yet universally resonant frameworks.

The research addresses such primary questions: (1) How are motifs associated with birth, life, death, and immortality distributed across Ukrainian and Estonian folk song corpora? (2) Can multilingual transformer models identify and map these motifs without manual annotation, and to what extent do cross-lingual embeddings capture culturally specific conceptualizations of the life cycle? (3) How can computational visualization of semantic transitions inform comparative studies of human existential experience in oral poetics?

The research is based on two main corpora: Ukrainian folk songs from the Podillia region and Estonian folk songs from Järvamaa, maintained by the Estonian Literary Museum. Together, these corpora provide a cross-linguistic platform for studying existential motifs in oral tradition.

The research applies recent transformer-based multilingual models for zero-shot semantic classification and embedding-based mapping. These models enable analysis without supervised training or parallel data, allowing exploration of conceptual structures directly from text.

Using mDeBERTa-v3-base-mnli-xnli (Laurer 2024) for zero-shot classification, each song or stanza receives probabilistic scores for the four conceptual categories (birth, life, death, immortality). This step identifies dominant existential themes and transitional areas where motifs overlap. Zero-shot inference is selected for its ability to generalize semantic meaning across languages without fine-tuning, enabling robust analysis of corpora.

For semantic embedding and similarity analysis, this study employs three multilingual transformer models: intfloat/multilingual-e5-large (Wang et al. 2024), sentence-transformers/stsb-xlm-r-multilingual (Reimers & Gurevych 2022), and paraphrase-multilingual-MiniLM-L12-v2 (Reimers & Gurevych 2019). They produce high-dimensional semantic representations of textual meaning for each song. To ensure comparability across models, embeddings are aligned and dimensionally reduced using Principal Component Analysis (PCA), after which embeddings are averaged to create a combined semantic representation.

Using cosine similarity and embedding distance metrics, the study traces semantic relationships among life-cycle stages (e.g., recurrent proximity between birth and immortality motifs). This allows for the identification of “semantic paths” corresponding to cyclical mythological thinking.

The methodology further incorporates visualization of the high-dimensional semantic space using UMAP (Uniform Manifold Approximation and Projection) to project embeddings into a two-dimensional space. This approach allows the mapping of existential motifs across songs and corpora, highlighting clusters and transitions between birth, life, death, and immortality. Interactive visualization using Plotly enables exploration of individual songs, stanzas, and their motif assignments, supporting both qualitative and quantitative interpretation of semantic patterns.

This research introduces a computational methodology for modelling mythological structures in multilingual folk song corpora. It demonstrates how zero-shot and embedding-based methods can extend computational folkloristics beyond theme extraction toward structural-semantic modelling of mythic cycles. By applying recent advances in multilingual language modelling, this research bridges traditional interpretive frameworks and computational cultural analysis, offering a scalable model for studying existential archetypes across oral traditions.

While grounded in Ukrainian and Estonian traditions, the methodology is adaptable to other languages and cultural corpora, contributing to digital preservation, comparative cultural analytics, and semantic modelling in the digital humanities.

References

1.Ariès, Philippe. (1981). The Hour of Our Death. Alfred A. Knopf.

2.Campbell, Joseph. (1949). The Hero with a Thousand Faces. Princeton University Press.

3.Eliade, Mircea. (1957). The Sacred and the Profane. RowohIt Taschenbuch Verlag GmbH.

4.Feifel, Herman. (1959). The Meaning of Death. McGraw-Hill.

5.Freud, Sigmund. (1955 [1920]). Beyond the Pleasure Principle. Translated by James Strachey. London: The Hogarth Press.

6.Heimann, Paula. (1952). Notes on the Theory of the Life and Death Instincts. In Melanie Klein, Paula Heimann, Susan Isaacs and Joan Riviere, Developments in Psycho-Analysis, pp. 321-337. London: Hogarth.

7.Laurer, Moritz. (2024). mDeBERTa-v3-base-mnli-xnli: A Multilingual DeBERTa Model for Cross-Lingual Natural Language Inference. Hugging Face Model Card.

8.Lévi-Strauss, Claude. (1966 [1962]). The Savage Mind. University of Chicago Press.

9.Reimers, Nils, & Gurevych, Iryna. (2019). Sentence-BERT: Sentence Embeddings Using Siamese BERT-Networks. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP). Association for Computational Linguistics.

10.Reimers, Nils, & Gurevych, Iryna. (2022). Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. Proceedings of EMNLP 2022: System Demonstrations. Association for Computational Linguistics.

11.Wang, Liang, Yang, Nan, Huang, Xiaolong, Yang, Linjun, Majumder, Rangan, & Wei, Furu. (2024). Multilingual E5 Text Embeddings: A Technical Report. arXiv. https://doi.org/10.48550/arXiv.2402.05672

Panel 1 — 13:30–15:30

13:30–15:30 **Digitizing Medicine from Below: Researcher-Driven Digitization for the History of Medicine**

Ylva Söderfeldt, Matts Lindström, Nils Hansson

13:30–15:30 *PANEL* **[14]**

Digitizing Medicine from Below: Researcher-Driven Digitization for the History of Medicine

Ylva Söderfeldt[1], Matts Lindström[1], Nils Hansson[2]

[1] *Uppsala university, Sweden*

[2] *Heinrich-Heine-University*

Keywords: *History of medicine, periodical print, dissertations, digitization, machine learning, digital archives*

1. Description

Four decades ago, Roy Porter made a famous call for a “medical history from below”, in which he pointed out the manifold ways that health and illness have been experienced and managed in everyday life, in a wide range of spaces besides the clinic, and by a multitude of people besides doctors (Porter 1985). However, with the introduction of digital methods for historical research, there is a risk that the field might revert back to focusing on canonical, mainstream sources, which usually come from the medical profession and powerful institutions (Thompson et al. 2016; Toon et al. 2016; Phillips et al. 2018; Coburn 2021; Milligan 2022). At the same time, digital methods carry a great potential to enhance the use and analysis of sources that are fragmented, diverse, and dispersed, and hence may become a powerful tool for a history of medicine that includes a wider variety of voices and perspectives. In this panel, we will present work from three projects that involve researcher-driven digitization of sources to the history of medicine and which explicitly aim to broaden the range of source materials available to digital analysis. In the process, the projects are addressing specific challenges that come with these types of materials: inconsistent and complex formatting, intermingling of a wide range of different content types, and complex issues of copyright, provenance, and data protection. This 105-minute panel introduces and invites attendees to discuss challenges and solutions in working with digital analysis of non-canonical sources to the history of medicine on the technical as well as analytical level.

The program for the panel is:

1. Introduction by panel chairs (5 minutes) 2. Paper presentations (45 minutes)

2.1. Matts Lindström and Sushruth Badri (Uppsala university, Sweden): Taming abundance: Machine Learning Applied to 200 Years of Medical Print

This paper presents the ongoing efforts of The Swedish Medical Periodicals project (SweMPer). The project aims to digitize, curate, and make accessible a vast but currently dispersed and fragmented collection of printed Swedish medical journals spanning over two centuries. When brought together as a searchable database, these periodicals will provide an invaluable resource for historians, researchers, and the general public. However, the heterogeneity of these documents, ranging from complex and varied layouts to different languages as well as degraded print quality poses significant challenges for digital processing and analysis. While such difficulties exist in any large-scale digitization effort, historical medical periodicals can present additional hurdles due to their specialized terminology, evolving typographic conventions, and domain-specific elements. We will present how the Swemper project leverages advanced machine learning techniques to overcome some of these challenges – despite the abundance and heterogeneity of the sources.

2.2. Ylva Söderfeldt and Gijs Aangenendt (Uppsala university, Sweden): The Patients’ View? How to Extract the Patient’s Voice from Digitized Patient Organization Magazines.

This paper presents the methodological challenges and solutions in ActDisease, a research project that seeks to establish the role played by patient organizations in medical knowledge generation, social policy, and healthcare practices in the 20th century. Within the project, we have created a database of digitized periodicals issued by patient organizations, and are developing methodological approaches to analyze it with a mix of digital and traditional methods. We will present how the ActDisease project works across disciplines to design methods that are tailored to historical research questions and how we as historians of medicine can integrate computational methods as part of our research process. It also addresses the question of what voices and actors that these non-canonical sources may be representing, and how this polyphony can be captured with digital means.

3.3.Nils Hansson, Thorsten Halling (Heinrich-Heine-University Düsseldorf, Germany), Hannah Ruschemeier (Osnabrück University, Germany), Miriam Albers (ZB MED Information Centre for Life Sciences, Cologne, Germany): Dissify: Digitizing European doctoral dissertations

This paper introduces preliminary results of the collaborative digitization project Dissify, which focuses on European medical dissertations from 1880 to 1950, funded by the German Research Foundation (2026-2028). Dissertations, though often invisible despite their value, are among the traditional forms of scholarship, making it essential to address their limited accessibility. In Germany, doctoral dissertations have been digitally archived through the German National Library’s DissOnline project since 1998. However, earlier dissertations remain fragmented across university libraries, often poorly cataloged. This project aims to make medical dissertations more discoverable and usable. It lays the foundation for a European online portal, Dissify: European Dissertation Cultures in Medicine, dedicated to the digitization and accessibility of medical dissertations.

3. Open Q & A-session (20 minutes)

4. General discussion (30 minutes)

The paper presentations will be followed by a discussion that addresses common themes across the presented projects and beyond:

- Copyright and the role of cultural heritage institutions: One major obstacle to digitization, which is especially pressing for modern material and for corpora outside of the mainstream, is copyright rules and their interpretation and implementation in cultural heritage institutions. We will discuss the way that these restrictions impact digitization and data management, in particular clashes with FAIR principles and the need for collaboration, as well as solutions for working with copyrighted material.
- Other non-canonical sources: We will discuss parallels between our digitization projects and other efforts to digitize and work with non-canonical sources, in particular potential synergy effects and the possibilities of developing best practices.
- Digitization bias: Making sources available digitally has the potential of amplifying their representation in historical and other research. We will discuss the risks involved in this and how researcher-driven digitization can reduce digitization bias.

Examples of questions for the discussion:

- Is there a tension between researcher-driven digitization efforts and mainstream digitization?
- Should memory institutions increase the availability of non-canonical sources, and if so, how?
- What biases might be introduced by researcher-driven digitization, and how do we make researcher-driven digitization projects useful beyond the confines of the specific project?

Acknowledgements

Communicating Medicine: Digitalisation of Swedish Medical Periodicals, 1781–2011 (SweMPer) is funded by Riksbankens Jubileumfond (IN22-0017)

ActDisease is funded by the European Union (ERC ActDisease, ERC-2021-STG 101040999). Views and opinions expressed are however those of the authors only and do not necessarily reflect those of the European Union or the European Research Council. Neither the European Union nor the granting authority can be held responsible for them.

Dissify is funded by the German Research Foundation (DFG, project number 565356721)

References

Coburn, Jon. 2021. “Defending the Digital: Awareness of Digital Selectivity in Historical Research Practice.” Journal of Librarianship and Information Science 53 (3): 398–410. https://doi.org/10.1177/0961000620918647.

Milligan, Ian. 2022. “The Transformation of Historical Research in the Digital Age.” Elements in Historical Theory and Practice, ahead of print, August. https://doi.org/10.1017/9781009026055.

Phillips, Christopher J., A. R. Ruis, Katherine Sorrels, et al. 2018. Viral Networks: Connecting Digital Humanities and Medical History. VT Publishing.

Porter, Roy. 1985. “The Patient’s View: Doing Medical History from Below.” Theory and Society 14 (2): 175–98.

Thompson, Paul, Riza Theresa Batista-Navarro, Georgios Kontonatsios, et al. 2016. “Text Mining the History of Medicine.” PLOS ONE 11 (1): e0144717. https://doi.org/10.1371/journal.pone.0144717.

Toon, Elizabeth, Carsten Timmermann, and Michael Worboys. 2016. “Text-Mining and the History of Medicine: Big Data, Big Questions?” Medical History 60 (2): 294–96. https://doi.org/10.1017/mdh.2016.18.

Poster Session — 16:15–18:00

16:15–18:00 **Claude in OCR of Historical Danish Hymnals 1740–1953**
Fedja Wierød Borčak

16:15–18:00 **Corpus of Danish novels 1855-1869**

Lasse Seistrup Holst, Thomas Hansen, Jens Bjerring-Hansen

16:15–18:00 **Making the Homosaurus Multilingual: A Community-Based Approach to Linked Open Data Translation**
Siska Humlesjö

16:15–18:00 **From Basement to Knowledge Graph: Bringing the Lars Dahle Card Catalogue to Life with AI**
Lars G Bagøien Johnsen, Jennifer Thøgersen, Live Rasmussen

16:15–18:00 **Navigating Semantic Abundance: Consensus Graph Clustering for Meaning Disambiguation in Coordination Networks**
Lars G Bagøien Johnsen

16:15–18:00 **Using 'Controlled Corpora' to Tame the Archived Web**
Christian Kaalund Kjeldsen, Helle Strandgaard Jensen

16:15–18:00 **Translocalis: Rediscovering Marginalized Readers' Letters in Finnish Newspapers, 1 886–1920s**
Heikki Kokko

16:15–18:00 **The Cautionary Tale of Women's Transatlantic Travel: A Study of Cartas de Llamada, a Forgotten Corpus**
THEODORA STAVROULA KORMA, OLGA ROJAS VALLE

16:15–18:00 **Visualizing (for) the Humanties**
Evelina Liliequist, Linnéa Tjernström, Maria Podkorytova

16:15–18:00 **"As Open as Possible, as Closed as Necessary": Balancing Openness, Sustainability, and Data Protection in Cooperative AI Infrastructure**
Christel Annemieke Romein, Melissa Terras, Andy Stauder, Florian Stauder, Michaela Prien

16:15–18:00 **Empowering humanities scholars with a modular digitisation pipeline**
David Rosson

16:15–18:00 **Svalbard in the Norwegian Press Imagination: Constructing an Arctic Nation**
Jana Sverdljuk, Lars Johnsen

16:15–18:00 **Preaching in Times of Crisis – A Large-Scale Text Study of Danish Sermons from Times of National Crisis**
Michael Mørch Thunbo

16:15–18:00 **Lost in Abundance, Found in Workflow: MagicTagger for Russian Tales – FAIR Knowledge Graph Export Enriched by a Folktale Type Classifier (Work-in-Progress Web Interface)**
Evgeniia Vdovichenko

16:15–18:00 **Semi-Automated Knowledge Graph Construction from Medieval Icelandic Sagas: Integrating CIDOC-CRMsoc, Shape Expressions, and Large Language Models**
Shintaro Yamada, Ikki Ohmukai

16:15–18:00 **A Digital Anchor: Cultivating Self-Leadership and Personal Agency in Youth through a Spiritual App**
Marcella Zoccoli, Klea Ziu

16:15–18:00 *POSTER & DEMO* **[15]**

Claude in OCR of Historical Danish Hymnals 1740–1953

Fedja Wierød Borčak
Aarhus University, Denmark

Keywords: *OCR, Claude API, digitization pipeline*

This poster presents an efficient workflow for digitizing historical Danish hymn collections and bible translations using Anthropic's Claude, developed at the Center for Digital Textual Heritage at Aarhus University. Traditional OCR approaches for historical materials face significant challenges, including Fraktur script, heteregeneous typographical layouts, extensive manual configuration, and labor-intensive post-OCR correction. Our pipeline addresses these challenges through three integrated phases: (1) OCR processing via Claude API with batch image input; (2) automated post-OCR correction including metadata extraction and text validation; and (3) database integration using a relational JSON structure linking metadata. The approach demonstrates several advantages: high accuracy with Fraktur

text, minimal configuration enabling rapid experimentation with different typographies, and the solution for traditional layout problems.

I will also include identified limitations and potential methodological problems. Non-deterministic outputs require robust quality assurance protocols; dependency on proprietary API raises concerns about long-term reproducibility; invisible errors necessitate systematic validation strategies. Preliminary results from digitizing Danish hymnals spanning the period 1740–1953 show promising accuracy on Fraktur text, significant reduction in manual correction time compared to traditional workflows, and successful processing of multiple distinct typographic layouts without reconfiguration.

The poster will present detailed workflow diagrams, comparative accuracy metrics, example outputs showing metadata extraction and text correction, and recommendations for implementing similar pipelines in other cultural heritage contexts. I discuss strategies for quality assurance and best practices for prompt engineering specific to historical texts.

16:15–18:00 *POSTER & DEMO* **[16]**

Corpus of Danish novels 1855-1869

Lasse Seistrup Holst[1], Thomas Hansen[1], Jens Bjerring-Hansen[2]

[1] *Danish Society for Language and Literature, Denmark*

[2] *Department of Nordic Studies and Linguistics, University of Copenhagen, Denmark*

Keywords: *corpora; Danish novels; Norwegian novels; metadata*

Abstract: Corpus of Danish novels 1855-1869

This paper introduces a new corpus of 252 Danish and Norwegian novels published in Denmark between 1855 and 1869. It is a part of a larger corpus of Danish and Norwegian texts called Tekstnet Text Corpus (TekstnetTC) that is currently under development. While most of the novels in the corpus fall outside the established canon and remain largely unknown today, they offer valuable insights into the period that is commonly referred to as the Danish Golden Age: what was read, which ideas circulated etc. The corpus thus represents a significant resource for research in literary, cultural, and intellectual history. In the paper, we discuss our principles of selection and describe the design and construction of the corpus. The corpus can be seen as a retrospective extension of the MeMo corpus (Bjerring-Hansen et al. 2022). In the paper, we briefly highlight the analytical, pragmatic, and infrastructural advantages of basing the corpus construction on original editions of the books, while also focusing specifically on novels. We describe how we identified and collected the novels using bibliographic resources (more specifically Danish bookseller catalogues), and how we used OCR tools to convert the novels into machine-readable texts. Furthermore, we present the extensive metadata set compiled for the novels and their authors which includes bibliographic information such as the novel's title, year of publication, number of pages, typeface, publisher, and sales price, as well as biographical information such as the author's name, pseudonym, and gender. Additionally, we provde examples of the aesthetic and cultural depth of the corpus by pointing to both emerging subgenres (such as urban fiction) and popular genres (such as crime fiction). Finally, we point to future uses and expansions of the corpus, both the corpus of novels and the larger Tekstnet Text Corpus.

References

Bjerring-Hansen, Jens, et al. "Mending Fractured Texts. A heuristic procedure for correcting OCR data." (2022). https://ceur-ws.org/Vol-3232/paper14.pdf

16:15–18:00 *POSTER & DEMO* **[17]**

Making the Homosaurus Multilingual: A Community-Based Approach to Linked Open Data Translation

Siska Humlesjö

GRIDH, Gothenburg University, Sweden

Keywords: *linked open data, controlled vocabularies, community translation, digital humanities, Homosaurus*

The *Homosaurus* is a Linked Open Data (LOD) thesaurus focusing on LGBTQ+ terminology, originally developed by IHLIA LGBT Heritage in the Netherlands in 1997 and currently maintained by the Digital Transgender Archive, Boston, USA. It contains more than 3,600 interlinked concepts describing LGBTQ+ identities, subcultural phenomena, cultural expressions, ethnic groups, intersex variations, and related topics. As a LOD resource, the *Homosaurus* is designed for integration across archives, libraries, and digital humanities projects, enabling consistent description and discovery of queer materials across collections and platforms.

A recent initiative aims to make the *Homosaurus* multilingual, with translations currently underway in Swedish, Dutch, German, French, and Japanese. The Swedish translation began with a subset of 700+ terms translated within the *Queerlit* project, which developed a LGBTQ+ thesaurus for fiction. In 2024, GRIDH (the Gothenburg Research Infrastructure in Digital Humanities) was invited by the *Homosaurus* board to produce a complete Swedish translation. This task has been carried out by a single research engineer and systems librarian—the author of this poster—but through a deliberately collaborative and community-based workflow.

The translation process has combined online consultations, in-person workshops, and outreach to organizations, activists, students, and researchers with specialized expertise in areas such as BDSM terminology, race and ethnicity, and medical language. This participatory model seeks to balance linguistic precision with inclusivity and community ownership.

As part of the digital workflow, ChatGPT was used as a support tool for pre-translation and consistency checking. Large language model outputs were not used as final translations but to accelerate preliminary drafts, identify inconsistent phrasing, and maintain terminological coherence across related entries. The poster will discuss how community translation practices and AI-assisted workflows intersect with the principles of Linked Open Data, and how these methods contribute to more inclusive, interoperable, and sustainable infrastructures for LGBTQ+ knowledge organization.

References

Homosaurus: An International LGBTQ+ Linked Data Vocabulary. Retrieved from https://homosaurus.org/about

Queerlit: Tesaurus. Retrieved from https://queerlit.dh.gu.se/om/tesaurus/

16:15–18:00 *SHORT PAPER* **[18]**

From Basement to Knowledge Graph: Bringing the Lars Dahle Card Catalogue to Life with AI

Lars G Bagøien Johnsen[1], Jennifer Thøgersen[2], Live Rasmussen[2]

[1] *National Library of Norway, Norway*

[2] *VID Specialized University Library*

Keywords: *digitization, special collections, language models, catalogue data, knowledge graphs*

Background and Motivation

Many libraries hold special collections accessible only through physical card catalogues, often stored in basements. One such case is the Lars Dahle Collection at VID Specialized University in Stavanger. Dahle's personal library, bequeathed in 1925, comprises around 5,000 volumes and offers a unique window into late-nineteenth- and early-twentieth-century missionary scholarship. This project leverages large language models (LLMs) to make this window more accessible to the world.

Project and Objectives

With support from the Eckbo's Endowment and in collaboration with the National Library of Norway, the overarching goal of the Dahle Library project was to increase the awareness of the Dahle collection. As part of this effort, we explored making the catalogue digitally accessible. After digitizing the cards, we tested out multiple methods to manipulate the data. This included cleaning and normalizing the data using OCR and several LLMs to prepare the catalogue cards to be connected to authority registers and bibliographic systems as well as form the basis for knowledge graphs.

Workflow and Data Processing

The processing was done in two rounds. The first round used OCR to extract the content of the cards and then used LLMs (Anthropic, Meta Llama) to structure the information; the second round shipped the images directly to the APIs of OpenAI. An important property of the cards that was learned from the first round, was that some cards had more than one entry on them. This was explicitly specified in the last round. The approximately 4000 cards yielded 4500 individual records this way.

The data contained numerous abbreviations, duplicates, and inconsistencies. The LLM was instructed to expand place abbreviations as well as modernizing, for example “Chra”. in card got both “Christania” as well as the modern form “Oslo”. The LLM also provided geolocation for the cities, enabling us to make a map of the publishing places.

Personal names were standardized, for example the cards often displayed the practice of distinguishing the first letter of names, like “L(ars) D(ahle)” which the LLM could normalize.

Authority registers were consulted to validate names, and fuzzy matching was tested to connect titles with entries in Alma and the National Bibliography. Multilingual title matching (Norwegian, English, German) of entries is planned for with the possibility of connecting the cards to digitally availability of full text Norwegian versions. For example, Lars Dahle’s own contributions in Norwegian were found using matching by author and title. We will show how Dahle’s own works can be analyzed using the geodata tools available at the DH-lab of the Norwegian National Library (Birkenes & Johnsen 2025).

Challenges and Model Limitations

Using an LLM directly on the images improved recall (up from 3000 to 4500 entries), but also introduced systematic errors, particularly with ambiguous abbreviations, and the abbreviations were not always consistent. For example, the abbreviation “Kra.” was usually expanded to “Kristiania”, but occasionally to “Krakow” instead. As such, prompt tuning and hybrid post-processing (human + automated) were necessary to maintain data reliability. The balance between automation and curation became a central methodological concern.

From Catalogue to Knowledge Graph

Once cleaned and structured, the catalogue emerges as a dataset in its own right. Each card—author, title, place, year—forms a node in a broader intellectual map. The data can be used to support linguistic, historical, and geographical analyses, revealing patterns in missionary scholarship. By treating the relationships of the cards as a graph, we can use clustering to study connectedness (Blondel et.al. 2008, Chakrabarti & Faloutsos 2012). A prototype JavaScript app now allows interactive exploration of this map and linkage to other corpora, which can access the API of the National Library to perform analysis as described in Birkenes et.al (2023).

Conclusion and Outlook

Dahle's Library demonstrates how digitization combined with LLM-based enrichment can transform archival catalogues from static aids into research resources. Future work includes full integration with Alma, visualization through knowledge graphs, and the extension of methods to similar special collections at VID and other Norwegian libraries.

References

Birkenes, M. B., & Johnsen, L. G. (2025). Corpus and the Bibliography: NB DH-LAB as an Infrastructure for Text and Metadata I J.-M. Hanssen & S. Furuseth (Red.), The Hermeneutics of Bibliographic Data and Cultural Metadata. Oslo: Notabene.

Blondel, V.D., Guillaume, J.-L., Lambiotte, R. & Lefebvre, E. (2008). "Fast unfolding of communities in large networks." Journal of Statistical Mechanics: Theory and Experiment, 2008(10), P10008.

Chakrabarti, D. and Faloutsos, C. (2012). Graph Mining. Morgan & Claypool Publishers.

16:15–18:00 *POSTER & DEMO* **[19]**

Navigating Semantic Abundance: Consensus Graph Clustering for Meaning Disambiguation in Coordination Networks

Lars G Bagøien Johnsen

National Library of Norway, Norway

Keywords: *clustering, semantic analysis, word2vec, collocations, coordination*

In this poster I want to describe how to integrate graph clustering methods that will overcome the partitioning effect of traditional clustering methods. These force each word into a single semantic or discursive category, hiding the polysemous nature of language. I will address the "lost in abundance" problem by proposing a consensus-based approach that combines overlapping clustering methods to reveal multiple semantic structures.

We analyze coordination networks ("X and Y" constructions) extracted from the National Library of Norway's digital collections and Google N-gram data for Norwegian, German, and English (Johnsen 2016, Breder Birkenes et al., 2015). Coordination networks provide semantically transparent relationships, as coordinated words typically share semantic features (Turney & Pantel, 2010). From these networks, we construct graphs where words are nodes and coordination relationships are weighted edges using pointwise mutual information. We then apply three complementary clustering methods: k-clique clustering, which identifies tightly connected word groups with high precision but misses peripheral terms (Chakrabarti & Faloutsos, 2012); Louvain community detection, which provides complete coverage but forces polysemous words into single categories (Blondel et al., 2008); and line graph clustering.

Line graph clustering, where edges become nodes and edge-to-edge relationships form new edges, naturally produces overlapping clusters when unpacked, allowing words to belong to multiple groups, suitable for semantic analysis. This technique, imported from physics and social network analysis into corpus linguistics, enables genuine polysemy representation. We demonstrate the approach with polysemous words across languages. Norwegian "kirsebær" (cherry) successfully separates into berry (jordbær, bringebær, blåbær) and wood/tree (eik, bjerk, furu) meanings through line graph clustering. Similarly, "pålegget" (definite form) distinguishes culinary (cheese, butter, bread) from legal/administrative (regulations, provisions, requirements) senses.

Our consensus methodology aggregates information from all three approaches, treating them as complementary "views" of semantic structure (Monti et al., 2003; Strehl & Ghosh, 2002). Words receive weighted membership scores in clusters based on votes from each method, similar to topic-word distributions in LDA but derived from graph topology rather than probabilistic modeling (Moisl, 2015). This multi-method triangulation reveals semantic nuances that single methods obscure, particularly for words with unequal frequency distributions across meanings. The approach extends naturally to word2vec (e.g. by recursive application of similarity resulting in a graph structure) similarity networks and document co-occurrence graphs for topic modeling, offering advantages over traditional partitioning methods by allowing documents and words to belong to multiple topics without stochastic complexity.

Evaluation combines structural and lexical comparison. We test whether consensus clustering separates dictionary-attested senses (e.g., polysemous forms such as *is*) more consistently than single partition-based methods. Cluster coherence is measured through intra-cluster PMI density and separation from competing clusters. Preliminary results indicate that overlapping graph structures preserve minority senses that would otherwise be absorbed into dominant partitions. A fuller comparative evaluation across graph methods will be presented in the proceedings version.

References

Blondel, V. D., Guillaume, J.-L., Lambiotte, R., & Lefebvre, E. (2008). Fast unfolding of communities in large networks. Journal of Statistical Mechanics: Theory and Experiment, 2008(10), P10008.

Breder Birkenes, M., Johnsen, L. G., Lindstad, A. M., & Ostad, J. (2015). From digital library to n-grams: NB N-gram. In Proceedings of the 20th Nordic Conference of Computational Linguistics (pp. 293-295). Linköping University Electronic Press.

Chakrabarti, D., & Faloutsos, C. (2012). Graph Mining. Morgan & Claypool Publishers.

Johnsen, L. G. (2016). Graph analysis of word networks. In International Symposium on Digital Humanities, 37–38.

Moisl, H. (2015). Cluster Analysis for Corpus Linguistics. De Gruyter Mouton.

Monti, S., Tamayo, P., Mesirov, J., & Golub, T. (2003). Consensus clustering: A resampling-based method for class discovery and visualization of gene expression microarray data. Machine Learning, 52(1-2), 91-118.

Strehl, A., & Ghosh, J. (2002). Cluster ensembles—a knowledge reuse framework for combining multiple partitions. Journal of Machine Learning Research, 3, 583-617.

Turney, P. D., & Pantel, P. (2010). From frequency to meaning: Vector space models of semantics. Journal of Artificial Intelligence Research, 37, 141-188.

16:15–18:00 *POSTER & DEMO* **[20]**

Using ‘Controlled Corpora’ to Tame the Archived Web

Christian Kaalund Kjeldsen, Helle Strandgaard Jensen
Aarhus University, Denmark

Keywords: *Archived Web, source criticism, childhood, Internet guidebooks*

The Internet has become a factor in many aspects of modern life worldwide over the past three decades. However, much contemporary history still ignores the archived web as a critical source for interpreting the recent past.[1],[2],[3]. There are many reasons for this. Several stem from the massive amount of data. Even when narrowed to a single URL, the volume of archived data from a single website can be overwhelming. Reading an old website up close in a playback view like the Internet Archive’s Wayback Machine also has several problems [4]. Most of all, however, finding out what the Web contained during a specific period is almost impossible, as archived Web pages aren’t easily searchable, and we don’t know how much of the live Web at the time was archived. If you want to search the Wayback Machine, you, for instance, have to know the name of the URL you are looking for, and you have to be lucky that it was archived during the period you are interested in analysing.

In the WEB CHILD project, we are interested in understanding how the introduction of the World Wide Web changed childhoods in South Korea, Denmark, and the United States 1995-2005. Naturally, this project relies heavily on archived web pages as one of its source types. In our poster presentation, we will demonstrate one approach we used to identify which websites were available to children during our period of interest, without knowing their URLs in advance.

Guidebooks that introduced children to the Web, offer an important insight into what adults believed children should explore online. They are an important starting point even if they represent the known problem that children use a lot of media beyond that intended for them [5], and thus do not capture many of the pages children would visit or make themselves.

To create a systematic overview of all available guidebooks published 1995-2005, we have searched libraries with legal deposit repositories for internet guidebooks for children, and any resources meant to aid adults in getting children online. Then we have registered any URLs they mention in a spreadsheet. The full spreadsheet thus provides an overview of all the websites we should be looking for in web archives, which together represent the web-sphere that adults believe children should be introduced to in the early era of the WWW.

This controlled and curated corpus furthermore enables us to apply principles from traditional source criticism, which is not otherwise immediately possible when working with archived web material. Knowing the authors and the context in which the books were produced allows us to contextualise their intentions regarding children’s web presence. This entry point to working with the archived web can be seen as working with controlled corpora; it is a controlled and systematic way of working with a subset of URLs, which can be traced back to a coherent set of ideas applied by the authors of the guidebooks, regardless of the specific URL content and availability.

References

Brügger, Niels. The Archived Web: Doing History in the Digital Age. The MIT Press, 2018.

Buckingham, David, Hannah Davies, Ken Jones, and Peter Kelley. Children’s Television in Britain: History, Discourse and Policy. BFI Publishing, 1999.

Milligan, Ian. History in the Age of Abundance? How the Web Is Transforming Historical Research. McGill-Queen’s University Press, 2019.

Schafer, Valérie, and Benjamin G. Thierry. “Web History in Context.” In The SAGE Handbook of Web History, edited by Niels Brügger and Ian Milligan. SAGE, 2018.

Winters, Jane. “Web Archives and (Digital) History: A Troubled Past and a Promising Future?” In The SAGE Handbook of Web History, edited by Niels Brügger and Ian Milligan. Sage, 2018.

16:15–18:00 *POSTER & DEMO* **[21]**

Translocalis: Rediscovering Marginalized Readers' Letters in Finnish Newspapers, 1886–1920s

Heikki Kokko
Tampere University, Finland

Keywords: *Readers' letters, Historical newspapers, Digital cultural heritage, Heritagization, Database*

The Translocalis project is a digital humanities initiative dedicated to uncovering and analyzing "local letters" (paikalliskirjeet) published in Finnish-language newspapers from the early 19th century through the 1920s. These readers' letters, often authored by anonymous or pseudonymous local correspondents, provided unique experiential insights into everyday life across Finland during its first modernization period. Despite their popularity and abundance in the press, local letters were marginalized by both contemporary elites and later historical research, largely due to their hybrid genre, limited accessibility, and perceived low cultural value. As a result, they have remained largely invisible in archives and scholarship, even though they represent the first substantial body of source material written by Finnish-speaking individuals in Finland.

The initial phase of the Translocalis project (1820–1885) involved the manual collection of 72,000 local letters from Finnish-language newspapers. Research assistants systematically reviewed every issue published up to 1885, creating a searchable database now available through the Finnish National Library's digital collections. This corpus has enabled new historical, linguistic, and cultural analyses of local experience and communication in Finland.

However, the rapid expansion of the Finnish-language press after 1885 made manual collection unfeasible. The current phase of the project seeks to expand Translocalis into the 1920s by leveraging digital methods and machine learning. The manually collected corpus now serves as training data for developing models capable of identifying similar texts in large-scale digitized newspaper corpora. The project aims to recover marginalized textual forms that are often overlooked by metadata and search algorithms, thereby making these voices accessible for further study.

Key elements of the poster include:

- A definition of the local letter phenomenon for an international audience unfamiliar with Finnish press.
- Criteria for identifying local letters, such as locative place names in titles, the presence of dates, pseudonyms, and expressions of local experience (e.g., "here," "in our parish").
- A methodology that combines manual annotation with machine learning to detect relevant texts in digitized newspaper corpora.
- An expansion plan to extend the database into the early 20th century, capturing the evolution of local discourse during Finland's modernization and political transformation.
- Shows how the recovery of local letters participates in heritagization and expands the scope of digital cultural heritage by bringing neglected, non-canonical materials back into view.

The project's relevance to DHNB 2026 is closely tied to the conference theme, "Lost in Abundance." Translocalis directly addresses the tension between digital abundance and invisibility by recovering texts that were once plentiful but are now digitally hidden. The project raises broader questions about genre, metadata, and the challenges of making non-canonical sources visible in the digital age.

Looking ahead, the poster outlines a roadmap for expanding Translocalis into the 1920s. This includes technical adaptations to new newspaper layouts and linguistic shifts, integration with other cultural heritage platforms, and potential applications in teaching, public history, and citizen science. By combining traditional scholarship with digital innovation, Translocalis seeks to illuminate the everyday voices of Finland's past and contribute to broader discussions about abundance, marginality, and digital recovery in the humanities.

References

Kokko, H. (2023). Translocalis Database. Digital Collections of the National Library of Finland. https://digi.kansalliskirjasto.fi/collections?id=742 (Accessed 27 October 2025).

Kokko, H. (2024). Village Gossips or Voice of the People? The Culture of Letters to the Press in the Grasp of Transnational Ideologies in Mid-1800s Finland. In J. Kortti & H. Kurvinen (Eds.), Mediated Ideologies: Nordic Views on the History of the Press and Media Cultures (pp. 3–20). Vernon Press.

Kokko, H. (2024). The Construction of Early Social Citizenship: The Lived Institution of Poor Relief in Mid-Nineteenth-Century Finland. In J. Annola, H. Lindberg & P. Markkola (Eds.), Lived Institutions as History of Experience. Palgrave Studies in the History of Experience. Palgrave Macmillan, Cham. https://doi.org/10.1007/978-3-031-38956-6_5

Kokko, H. (2022). From Local to Translocal Experience: The Nationwide Culture of Letters to the Press in Mid-1800s Finland. Media History, 28(2), 181–198. https://doi.org/10.1080/13688804.2021.1961575

Kokko, H. (2021). Temporalization of Experiencing – First-Hand Experience of the Nation in Mid-Nineteenth Century Finland. In V. Kivimäki, S. Suodenjoki & T. Vahtikari (Eds.), Lived Nation as the History of Experiences and Emotions in Finland, 1800–2000 (pp. 109–134). Palgrave Macmillan. https://doi.org/10.1007/978-3-030-69882-9_5

16:15–18:00 *POSTER & DEMO* **[22]**

The Cautionary Tale of Women's Transatlantic Travel: A Study of Cartas de Llamada, a Forgotten Corpus

THEODORA STAVROULA KORMA[1,2,3,4], OLGA ROJAS VALLE[1]

[1] *Rijkuniversiteit Groningen, The Netherlands*
[2] *Linnaeus University, Sweden*
[3] *Stockholm University, Sweden*
[4] *Vienna Contemporary Art Space, Austria*

Keywords: *Digital humanities, underexplored archives, women's travel, Spanish colonial era, transatlantic migration, Cartas de Llamada, Old Spanish, GIS mapping, marginalised voices*

This project investigates Cartas de Llamada, a largely overlooked collection of letters sent from Spanish colonies to Spain between the 15th and 19th centuries. These letters were both personal and official, allowing senders to invite relatives to the colonies and maintain contact with family, while the Spanish Crown used them to regulate transatlantic migration. Although preserved in the Archives of Seville, these letters were largely disregarded, offering a rich source for understanding travel, social conditions, and the roles of Indigenous peoples in colonial life. From this collection, of particular interest are the recommendations aimed at women travelling across the Atlantic, revealing everyday challenges, social expectations, and practical advice rarely captured in historical records. The corpus also includes voices from all social classes, including commoners often ignored in traditional histories, providing a fuller picture of colonial society.

The methodology and digital approaches we will follow include translation and textual analysis, with selected letters translated from Old Spanish to English, making the corpus accessible for English-speaking researchers. The letters are also divided into key themes that have been obtained through topic modeling (NLP). Using previously prepared data, the letters will also be mapped with ArcGIS to visualize transatlantic travel routes, migration patterns, and points of settlement. The project critically examines biases in archival preservation and historical narratives, amplifying marginalized voices that have often been overlooked. By combining historical research, literary analysis, and digital humanities methods, it uncovers underexplored cultural and social material.

Cartas de Llamada fits the "Lost in Abundance" theme by revealing a rich but overlooked corpus within well-known archives in Spain. Focusing on women's travel and voices from all social strata, the project demonstrates how digital humanities can recover and analyse neglected historical sources. Mapping, translation, and textual analysis bring forgotten material to light, offering new insights into the social, cultural, and gendered dimensions of the Spanish colonial world.

References

DeBats, D. A., & Gregory, I. N. (2011). Introduction to historical GIS and the study of urban history. Social Science History, 35(4), 455–463. https://doi.org/10.1215/01455532-1381814

El-Sieedy, A., Abuzekry, T., & Al-Menshawy, A. (2021). The difference between the concept of space and of place in urban science. The Egyptian International Journal of Engineering Sciences and Technology, 35(1), 1–7. https://doi.org/10.21608/eijest.2021.57319.1038

Edelstein, D., Findlen, P., Ceserani, G., Winterer, C., & Coleman, N. (2017). Historical research in a Digital Age: Reflections from the Mapping the Republic of Letters project. The American Historical Review, 122(2), 400–424. https://doi.org/10.1093/ahr/122.2.400

Geddes, A., & Gregory, I. (2014). From historical GIS to spatial humanities: An evolving literature. In I. N. Gregory & A. Geddes (Eds.), Towards Spatial Humanities: Historical GIS and Spatial History (pp. 186-202). Indiana University Press.

Goodchild, M. F. (1992). Geographical information science. International Journal of Geographical Information Systems, 6(1), 31–45. https://doi.org/10.1080/02693799208901893

Goodchild, M. F. (2010). Geographic Information Systems. In B. Gomez & J. P. Jones III (Eds.), Research Methods in Geography: A Critical Introduction (pp. 376–392). Wiley-Blackwell.

Gregory, I. N., Donaldson, C., Murrieta-Flores, P., & Rayson, P. (2015). Geoparsing, GIS and textual analysis: Current developments in Spatial Humanities research. International Journal of Humanities and Arts Computing, 9(1), 1-14. https://doi.org/10.3366/ijhac.2015.0135

Gregory, I., & Ell, P. S. (2012). Historical GIS: Technologies, Methodologies, and Scholarship. Cambridge University Press.

Gregory, I., Bushell, S., & Cooper, D. (2015-2018). Geospatial Innovation in the Digital Humanities: A Deep Map of the English Lake District. LancasterUniversity.

Herrmann, J. B. (2017). In a test bed with Kafka: Introducing a mixed-method approach to digital stylistics. Digital Humanities Quarterly, 11(4). https://www.digitalhumanities.org/dhq/vol/11/4/000341/000341.html

Manovich, L. (2017). Cultural analytics, social computing and digital humanities. In M. T. Schäfer & K. van Es (Eds.), The datafied society: Studying culture through data (pp. 55–68). Amsterdam University Press. https://doi.org/10.25969/mediarep/12514

16:15–18:00 *POSTER & DEMO* **[23]**

Visualizing (for) the Humanties

Evelina Liliequist, Linnéa Tjernström, Maria Podkorytova

Umeå University, Sweden

Keywords: *Visualization, methods, infrastructure, projects, showcase*

Scientific visualization enhances our ability to understand complex data across a wide range of disciplines and enables researchers to communicate their findings in clear, accessible, and innovative ways. As academia becomes increasingly multidisciplinary, visualization also serves as a shared language that bridges methodological and conceptual gaps between fields. Within the humanities, visualization not only complements established research practices but expands them. It can assist Humanities scholars to uncover hidden patterns, trace cultural and historical developments, and communicate their findings through visual storytelling.

InfraVis, Sweden's national infrastructure for data visualization and analysis, provides this kind of specialized support to researchers at universities across the country. Through expert guidance, advanced technologies, and state-of-the-art visualization tools, InfraVis helps scholars transform raw, heterogeneous data into meaningful and interpretable forms. Our poster presentation highlights several examples of how data visualization can enrich humanities research.

One such example is NordPow (Gelfgren & Tjomsland 2024), a Nordic project that brings together data on churches and prayer houses across Sweden and Norway. Beyond its value for research, it also helps document and preserve an important religious heritage. The project focused on collecting, harmonizing, and merging national datasets into one coherent and durable resource. The data was then visualized in ArcGIS and turned into an interactive web application using ArcGIS Experience Builder. Users can explore the material through filters for religious affiliation and time periods, as well as an interactive timeline that shows how the landscape changes over time. Learn more as we showcase this and some more examples that demonstrate how InfraVis – with our team of over 50 visualization experts, across nine universities – can support humanities research by helping researchers explore, interpret, and communicate data in new and meaningful ways.

References

https://infravis.se

16:15–18:00 *POSTER & DEMO* **[24]**

"As Open as Possible, as Closed as Necessary": Balancing Openness, Sustainability, and Data Protection in Cooperative AI Infrastructure

Christel Annemieke Romein[1,2], Melissa Terras[1,3], Andy Stauder[1], Florian Stauder[1], Michaela Prien[1]

[1] *READ-COOP SCE, Austria*

[2] *University of Twente, Enschede, the Netherlands*

[3] *University of Edinburgh, Edinburgh, Scotland*

Keywords: *Sustainability, Automatic Text Recognition, Cooperative infrastructure, Data protection, Open Science*

READ-COOP, established in 2019 as a European Cooperative Society, maintains and develops Transkribus, a platform for Automated Text Recognition of historical documents (Romein et al. 2025). Since its inception, the cooperative has navigated a critical tension for responsible ML/AI providers: balancing the imperatives of development, openness, and accessibility with the practical necessities of data protection, intellectual property preservation, and financial sustainability. This poster examines READ-COOP's approach to this challenge through the principle of being "as open as possible, as closed as necessary": a phrase originally concerned with GDPR (Landi et al. 2020), and now being used within the library sector to describe the boundaries that need to be put in place to ensure longevity, in a time of predatory scraping for Generative-AI (Bryant 2025).

The cooperative faces several constraints requiring selective closure of its infrastructure. GDPR compliance mandates careful stewardship of personal data and user information. Additionally, the business model relies on protecting the proprietary algorithms and extensive training datasets that form READ-COOP's core intellectual property—assets developed over years with significant public investment. Without this protection, larger technology providers could appropriate these resources, threatening the cooperative's viability and its ability to serve its community. The extractive practices of major AI platforms, including the illegal scraping of cultural heritage content, make complete openness untenable for a sustainable infrastructure serving the library, archive, and museum sectors.

However, READ-COOP remains committed to maximum accessibility through multiple mechanisms. The platform offers free monthly credits to enable trial use and support hobbyist researchers. A comprehensive Scholarship Programme provides free access to students from 73 countries, having supported 336 individuals with over 1 million free credits by October 2024 (Nockels, Gooding, and Terras, 2025). Users retain full ownership of their data and can export all materials in open formats. The cooperative publishes 300 licensed AI models for internal community use, and encourages members to share training data through repositories such as HTR-United. Transkribus Sites enables researchers to publish digital editions without institutional infrastructure or coding expertise.

The cooperative structure itself embodies openness through democratic governance, enabling members to provide meaningful input into platform development and strategic decisions (Terras et al. 2025). Monthly meetings and annual conferences foster transparent dialogue between users and developers. This approach demonstrates that ethical AI governance requires not merely technical openness but also structural mechanisms that ensure community participation, accountability, and shared ownership.

READ-COOP's experience suggests that responsible ML/AI development necessitates pragmatic balance rather than absolutist positions. By prioritising community needs, transparent governance, and sustainable business practices over profit, the cooperative model offers an alternative framework for AI infrastructure—one that protects both the technology and the community it serves, demonstrating that closure and openness need not be oppositional but can be strategically calibrated to ensure long-term viability and broad accessibility. We suggest that this may be the reality of openness in the current digital humanities age: reframing our approach to ensure our digital assets are not stripped from us by unscrupulous AI providers.

References

R. Bryant (2025), Open research as a strategic priority: Insights from an OCLC RLP leadership roundtable (Blogpost)

A. Landi, M. Thompson, V. Giannuzzi, F. Bonifazi, I. Labastida, L.O.B. da Silva Santos, and M. Roos (2020) “The “A” of FAIR–as open as possible, as closed as necessary.” in: Data Intelligence 2020; 2 (1-2): 47–55

J. Nockels, P. Gooding, and M. Terras (2025), “Are Digital Humanities platforms facilitating sufficient diversity in research? A study of the Transkribus Scholarship Programme”, in: Digital Scholarship in the Humanities, 40 (Supplement 1), pp. 146-165.

C.A. Romein, M. Terras, A. Stauder, B. Anzinger, F. Stauder (Mai 2025), Praeteritum transcriptum. A Transkribus Tribute: Celebrating our First Five Years as a Cooperative (2019-2024), READ COOP SCE, Innsbruck. https://zenodo.org/records/15308678 ISBN: 9798284999653

M. Terras, B. Anzinger, P. Gooding et al. (2025) “The artificial intelligence cooperative: READ-COOP, Transkribus, and the benefits of shared community infrastructure for automated text recognition” [version 2; peer review: 1 approved, 1 not approved]. Open Research Europe, 5:16 (https://doi.org/10.12688/openreseurope.18747.2)

16:15–18:00 *POSTER & DEMO* **[25]**

Empowering humanities scholars with a modular digitisation pipeline

David Rosson

University of Helsinki, Finland

Keywords: *digitisation, historical documents, research infrastructure, OCR, text reuse*

Digital humanities and intellectual history rely on large-scale digitisation of historical sources. This is a highly technical process with barriers and quality issues that limit its access and applicability for scholars from analogue traditions. This poster, accompanied by software demonstrations, presents a workflow to address the current limits. The workflow (or digitisation “pipeline”) is a set of methodology guidelines and documentation, supported by open source software packages serving as example implementations.

The pipeline covers the end-to-end cycle: from warehousing of page images, to geometric image pre-processing, to document layout analysis, to OCR, to automated or editorial post-correction, to machine-readable (TEI-like) digital objects, to downstream use case such as text overlap detection, semantic search, and automated collation, supported by web-based user interfaces for scholarly exploration.

The functional features parallel those of commercial tools (e.g. Transkribus), though the goal is to enhance transparency, control, and flexibility via a computing-focused and modular approach versus pre-packaged products. Each module incorporates state-of-the-art, openly available techniques, e.g. YOLO for layout detection, Tesseract or neural methods for OCR, passim for alignment, transformers for text correction, bioinformatic and edit-distance algorithms for collation and reuse detection to produce interoperable datasets. As the frontier of technology develops, each module can be swapped to a more performant alternative.

The pipeline’s prototypes leverage support from CSC, a cloud computing provider (comparable to AWS, Google Cloud, Azure) owned by the Finnish state and universities, which offers subsidised computing resources (including hosting, storage, data processing) to researchers. By adopting a config-driven, “workflow as code” approach, the processing steps are run through container orchestration (Kubernetes), making each modular task repeatable with stable dependencies and the pipeline adaptable for a variety of datasets and projects.

References

Rosson, D., Mäkelä, E., Vaara, V., Mahadevan, A., Ryan, Y., & Tolonen, M. (2023). Reception Reader: Exploring Text Reuse in Early Modern British Publications. Journal of Open Humanities Data, 9(5), 1–11. https://doi.org/10.5334/johd.101

Péter, R. (2025). Uncovering Hidden Influences: The Reception Reader as a Tool for Intellectual Historians. Global Intellectual History, 1–13. https://doi.org/10.1080/23801883.2025.2474486

16:15–18:00 *SHORT PAPER* **[26]**

Svalbard in the Norwegian Press Imagination: Constructing an Arctic Nation

Jana Sverdljuk, Lars Johnsen

National Library of Norway, Norway

Keywords: *Digital text analysis, Corpus linguistics, Svalbard, Norwegian press, Arctic identity, Sovereignty*

The presentation traces how Norwegian newspapers imagined Svalbard as part of the nation. By applying digital humanities methods developed at the National Library of Norway (https://www.nb.no/dh-lab/apper/), in combination with discourse analysis (Fairclough 1992; Foucault 1972), it identifies the main spatial imaginaries and communicative registers in press coverage of Svalbard. As the analysis shows, these are circumpolar, mainland-territorial and regional Northern Norwegian imaginaries. Especially in the period from the early 1920ies to 1940ies, newspapers primarily include short operational and practical reports on fishing, coal mining infrastructure updates, meteorological bulletins and scientific notices. This may be understood as the reproduction of national belonging through seemingly ordinary informational practices.

16:15–18:00 *POSTER & DEMO* **[27]**

Preaching in Times of Crisis – A Large-Scale Text Study of Danish Sermons from Times of National Crisis

Michael Mørch Thunbo

Aarhus University, Denmark

Keywords: *quantitative text analysis; church history; homiletics, crisis rhetoric.*

Danish church history has richly narrated the nineteenth and twentieth centuries, yet much of this work remains qualitative and episodic. This PhD project offers a diachronic, large-scale view of how sermons - one of the most influential genres of religious communication - frame crisis, comfort, agency, and national identity. By assembling and computationally analyzing a curated corpus of Danish sermons, I map rhetorical and affective patterns across time and theological traditions to clarify how crisis is interpreted and mediated in homiletical practice. I ask: (1) How were specific national or global crises addressed, and did responses vary across time or tradition? (2) Which textual and thematic features define a “crisis sermon,” and how can that genre be operationalized for analysis? (3) What theological interpretations, coping mechanisms, and forms of collective identity are articulated in moments of upheaval?

The dataset comprises ~1,800 digitized sermons (1853–1963) from the Royal Danish Library (published collections, newspapers, and church periodicals), sampled around seven crisis windows across epidemic, war, economic downturn, occupation, and Cold War anxiety, plus a small comparative subset from 2020–2021 (COVID-19). Sermons are stored as cleaned .txt files and linked to metadata on preacher, date, crisis period, liturgical context, and location, enabling longitudinal and tradition-sensitive comparison. Because the corpus spans more than a century of Danish prose, diachronic linguistic variation is treated as a methodological and interpretive condition through normalization and period-sensitive robustness checks.

Currently halfway through the four-year dissertation, I have completed data collection and am developing the project as a series of articles (two in press; one to be finalized in March). Article 1 establishes tradition-specific theological vocabularies (1914–1963) using frequency analysis and pointwise mutual information (PMI), showing that Inner Mission and Grundtvigian sermons diverge in theological anthropology and discourses of sin (legal/forensic vs. existential/relational), shaping how agency, culpability, and consolation are articulated. Article 2 uses topic modeling, Danish sentiment analysis (SENTIDA), and change-point detection to track tonal and thematic inflections during the occupation (1940–1945); despite overall stability, clear shifts cluster around the 1942 Telegram Crisis and the 1943 government collapse, with traditions responding in distinct registers of nation, fear, and responsibility. Article 3 conducts Quantitative Close Reading of 295 sermons across three crises, showing that crisis-explicit sermons recurrently construe crisis as a shared condition, intensify fear/vulnerability language, exhort resilience and social cohesion, activate communal identity frames, and shift from lectionary-centred exposition toward contextual meaning-making. Interpreting these patterns through Victor Turner’s account of liminality and redress, I theorize crisis sermons as threshold rhetoric that performs a redressive cycle (diagnosis → reframing → reintegration).

Methods and tools include DaCy for Danish NLP and custom reproducible Python scripts for corpus curation, modeling, and visualization. The project contributes a metadata-rich sermon corpus, an

operationalized concept of the “crisis sermon,” and quantitative analyses that link rhetorical change to questions of Danish religious culture and sense-making under disruption. It also bridges large-scale distant reading with historically informed close reading, engaging Digital Humanities debates on scale, interpretability, and disciplinary integration.

References

Selected:

Agersnap, Anne. Collective Testimonies to Christianity and Time: A Collection and Large-Scale Text Study of 11,955 Danish Sermons from 2011–2016. PhD diss., Aarhus Universitet, 2021.

Ammerman, Nancy Tatom. Sacred Stories, Spiritual Tribes: Finding Religion in Everyday Life. Oxford: Oxford University Press, 2013.

Arendt, Hannah. Between Past and Future: Eight Exercises in Political Thought. London and New York: Penguin Books, 1993.

Berry, David M., ed. Understanding Digital Humanities. Houndmills and New York: Palgrave Macmillan, 2012.

Berry, David M. “Introduction: Understanding Digital Humanities.” In Understanding Digital Humanities, edited by David M. Berry, 1–20. Houndmills and New York: Palgrave Macmillan, 2012.

Collins, Randall. Interaction Ritual Chains. Princeton: Princeton University Press, 2014.

Daiber, Karl-Fritz. Predigen und Hören: Ergebnisse einer Gottesdienstbefragung. Munich: Kaiser, 1980.

Dixon, Dan. “Analysis Tool or Research Methodology: Is There an Epistemology for Patterns?” In Understanding Digital Humanities, edited by David M. Berry, 191–209. Houndmills and New York: Palgrave Macmillan, 2012.

Durkheim, Émile. The Elementary Forms of Religion. Translated by Carol Cosman. Oxford: Oxford University Press, 2001.

Edwards, O. C. A History of Preaching. Nashville: Abingdon Press, 2004.

Engemann, Wilfried. Homiletics: Principles and Patterns of Reasoning. 2nd ed. Berlin and Boston: De Gruyter, 2019.

Evans, Leighton, and Sian Rees. “An Interpretation of Digital Humanities.” In Understanding Digital Humanities, edited by David M. Berry, 21–41. Houndmills and New York: Palgrave Macmillan, 2012.

Franks, Anne, and John Meteyard. “Liminality: The Transforming Grace of In-Between Places.” Journal of Pastoral Care & Counseling 61, no. 3 (2007): 215–22.

Glenthøj, Jørgen. Kirkelige dokumenter fra besættelsestiden: Officielle og uofficielle hyrdebreve, bekendelser og erklæringer fra den danske kirke, samt et tillæg fra den norske kirke og den tyske bekendelseskirke. Sabro, 1985.

Heimbrock, Hans-Günter. “Practical Theology as Empirical Theology.” International Journal of Practical Theology 14, no. 2 (2011): 153–70.

Jacobsen, Erik Thostrup. Som om intet var hændt: Den danske folkekirke under besættelsen. Copenhagen: Universitetsforlag, 1991.

Lorensen, Marlene Ringgaard. Dialogical Preaching: Bakhtin, Otherness and Homiletics. Göttingen: Vandenhoeck & Ruprecht, 2013.

Lundgren, Linnea, and Linnea Jensdotter. “Studying Religious Change: Combining Close and Distant Reading in the Field of Sociology of Religion.” Nordic Journal of Religion and Society 35, no. 2 (2022): 111–24.

McClure, John S. Other-Wise Preaching: A Postmodern Ethic for Homiletics. St. Louis: Chalice Press, 2001.

Mills, C. Wright. The Sociological Imagination. Oxford: Oxford University Press, 1959.

Moretti, Franco. Graphs, Maps, Trees: Abstract Models for a Literary History. London and New York: Verso, 2005.

Piotrowski, Michael. “Ain’t No Way Around It: Why We Need to Be Clear About What We Mean by ‘Digital Humanities.’” In Wozu Digitale Geisteswissenschaften? Innovationen, Revisionen, Binnenkonflikt, edited by M. Hober, S. Krämer, and C. Pias. Lüneburg: Leuphana University, 2019.

Rieder, Bernhard, and Theo Röhle. “Digital Methods: Five Challenges.” In Understanding Digital Humanities, edited by David M. Berry, 67–84. Houndmills and New York: Palgrave Macmillan, 2012.

Roslyng-Jensen, Palle. Danskerne og besættelsen: Holdninger og meninger 1939–1945. Copenhagen: Gads Forlag, 2007.

Sandbæk, Harald, and N. J. Rald, eds. Den danske Kirke under Besættelsen. Copenhagen: H. Hirschsprung, 1945.

Sherratt, Tim. “Hacking Heritage: Understanding the Limits of Online Access.” In The Routledge International Handbook of New Digital Practices in Galleries, Libraries, Archives, Museums, and Heritage Sites, 116–30. London: Routledge, 2019.

Sterne, Jonathan. “The Example: Some Historical Considerations.” In Between Humanities and the Digital, edited by Patrick Svenson and David Theo Goldberg, 17–33. Cambridge, MA, and London: MIT Press, 2015.

Taylor, Charles. A Secular Age. Cambridge, MA, and London: Harvard University Press, 2009.

Thomassen, Bjørn. Liminality and the Modern: Living Through the In-Between. London and New York: Routledge, 2016.

Thunbo, Michael Mørch, and Kristoffer Laigaard Nielbo. ‘En Metodisk Invitation: Komputationel Analyse Af Forkyndelsen i Indre Mission Og Den Grundtvigske Bevægelse (1914-1963)’. In Nordisk Väckelse-Forskning - Trender Och Möjligheter, edited by Carola Nordbäck. Svenska Kyrkans Forskningsserie (in Press). Stockholm, 2026.

Thunbo, Michael Mørch. ‘Preaching in Occupied Land: Crisis and Sermons in Wartime Denmark (1940–1945)’. Studia Theologica - Nordic Journal of Theology (in press), 2026. doi:10.1080/0039338X.2026.2622675.

Tisdale, Leonora Tubbs. Preaching as Local Theology and Folk Art. Minneapolis: Fortress Press, 1997.

Underwood, Ted. “Distant Reading and Recent Intellectual History.” In Debates in the Digital Humanities 2016, edited by Matthew K. Gold and Lauren F. Klein, 530–33. Minneapolis: University of Minnesota Press, 2016.

Turner, Victor. The Ritual Process: Structure and Anti-Structure. London and New York: Routledge, 2017.

Veleski, Stefan. “Nothing New Under the Sun? Computational Humanities and the Methodology of History.” In Proceedings of the Workshop on Computational Humanities Research, edited by Folgert Karsdorp, Barbara McGillivray, Adina Nerghes, and Melvin Wevers, 171–81. Amsterdam, 2020.

Wohlrab-Sahr, Monika, and Marian Burchardt. “Multiple Secularities: Toward a Cultural Sociology of Secular Modernities.” Social Compass 59, no. 2 (2012): 282–99.

16:15–18:00 *POSTER & DEMO* **[28]**

Lost in Abundance, Found in Workflow: MagicTagger for Russian Tales – FAIR Knowledge Graph Export Enriched by a Folktale Type Classifier (Work-in-Progress Web Interface)

Evgeniia Vdovichenko

University of Bologna, Italy

Keywords: *computational folkloristics; ATU; FAIR; Docker; UX for scholars; Russian folktales; LOD*

MagicTagger is a pilot prototype which deals with an “abundance problem” in folklore studies: even with digitised archives, the diversity and inconsistent descriptions of their contents make study hard, as the work relies on manual processes and requires specialist knowledge. The system is based on Russian folktales from the Estonian Folklore Archives – part of the Estonian Literary Museum, Tartu – and concentrates on tales of magic, the most common genre within the Russian holdings.

Our principal aim is to create, implement and assess a knowledge-management system which views the archive as a defined knowledge domain, rather than a set of documents. Archive records are presented as a knowledge graph which understands the origin of the material, and is intended for machine use, in accordance with FAIR principles. Folktale type is the key element for comparing the tales; it enables discovery, faceted exploration and comparative study. We employ the internationally recognised Aarne–Thompson–Uther (ATU) Index to align with current scholarly practice. Currently, comparison is limited to tales in Russian; finding tales by type is used to link and compare texts in the same language – though we are considering, for future development, comparing folktales across cultures using our knowledge graph.

Technically, MagicTagger has a web interface, presenting visualisations of the folktale collection from the Estonian Folklore Archive, the collection’s knowledge graph, and a machine-learning type classifier, which proposes the three most likely ATU types for a tale from outside the archive, and incorporates these suggestions into the metadata for subsequent use which you could export as an RDF.

The system’s outputs prioritise re-use and validation. It generates RDF and JSON-LD files of archive records and type assignments, with machine-readable provenance – therefore showing model versions, settings, times and checksums. Model performance is indicated by SHACL validation and a suite of SPARQL queries which enact the main knowledge-management tasks: discovery, comparison, control, and re-use. A first iteration of the web interface allows discovery–filtering–comparison, and export for further investigation. The system also classifies texts not yet categorised in the archive.

The project operates under the assumption of controlled access, as dictated by archive regulations: manuscripts remain subject to rights and privacy restrictions, and the outputs respect these. Within these constraints, MagicTagger demonstrates how a small, controlled, and provenance-aware system can enhance the ease with which material can be compared and reused, given the practical challenges of labour, access and institutional structures. The poster will demonstrate the system built, sample outputs, locating material in the graph by query, exports which reveal provenance, and screenshots of the interface.

References

1. Abello, J., Broadwell, P., & Tangherlini, T. R. (2012). Computational folkloristics. Communications of the ACM, 55(7), 60–70.
2. Afanas'ev, A. N. (1984–1985). Narodnye russkie skazki (Vols. 1–3). L. G. Barag & N. V. Novikov (Eds.); E. V. Pomerantseva & K. V. Chistov (Series Eds.). Moscow: Nauka. (Series: Literaturnye pamiatniki).
3. Cornell Research Data Management Service Group. (n.d.). Sharing and reusing data: FAIR. Retrieved October 28, 2025, from https://data.research.cornell.edu/data-management/sharing/fair/
4. Dalkir, K. (2013). Knowledge Management in Theory and Practice (0 ed.). Routledge. https://doi.org/10.4324/9780080547367
5. Danish Folklore Nexus. (n.d.). https://scando.ist.berkeley.edu/folklorenexus/
6. Declerck, T., Aman, A., et al. (2017). Multilingual Ontologies for the Representation and Processing of Folktales. In Proceedings of the Workshop on Language Technology for Digital Humanities in Central and (South-)Eastern Europe (pp. 20–23). https://doi.org/10.26615/978-954-452-046-5_003
7. Ilyefalvi, E. (2018). The Theoretical, Methodological and Technical Issues of Digital Folklore Databases and Computational Folkloristics. Acta Ethnographica Hungarica, 63(1), 209–258. https://doi.org/10.1556/022.2018.63.1.11
8. Järv, R., & Sarv, M. (2013). Estonian Folklore Archives. Oral Tradition, 28(2). https://doi.org/10.1353/ort.2013.0022
9. Li, Y. (n.d.). Folk Tales from Diverse Cultures: Digital Analysis of Content using Natural Language Processing [Manuscript].
10. Loss, E., Guernaccini, F., & Carassai, M. (n.d.). From Manuscript to Metadata: Experiments on Handwritten Text Recognition, Tagging and Importation for the Memoriali series (1265–1452) [Conference paper / project report].
11. Nguyen, D., Trieschnigg, D., & Theune, M. (n.d.). Folktale classification using learning to rank [Manuscript / proceedings paper].
12. Peroni, S., Tomasi, F., & Vitali, F. (2013). The Aggregation of Heterogeneous Metadata in Web-Based Cultural Heritage Collections: A Case Study. International Journal of Web Engineering and Technology, 8(4), 412. https://doi.org/10.1504/IJWET.2013.059107
13. Tangherlini, T. R. (n.d.). Challenges for a Computational Folkloristics.
14. Uther, H.-J. (2004). The Types of International Folktales: A Classification and Bibliography. Part I: Animal Tales, Tales of Magic, Religious Tales, and Realistic Tales, with an Introduction. Helsinki: Suomalainen Tiedeakatemia (FF Communications).
15. Uther, H.-J. (2011). The Types of International Folktales. Vol. 1: Animal Tales, Tales of Magic, Religious Tales, and Realistic Tales, with an Introduction (2nd print.). FF Communications / Ed. for the Folklore Fellows; Vol. 284 (= Vol. 133). Suomalainen Tiedeakatemia.
16. Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., et al. (2016). The FAIR Guiding Principles for Scientific Data Management and Stewardship. Scientific Data, 3(1), 160018. https://doi.org/10.1038/sdata.2016.18
17. WossiDiA. (n.d.). https://www.wossidia.de/#was
18. Zamesin.ru. (n.d.). About Jobs to Be Done (ProductHowTo). Retrieved October 28, 2025, from https://zamesin.ru/producthowto/book/about-jobs-to-be-done/

16:15–18:00 *POSTER & DEMO* **[29]**

Semi-Automated Knowledge Graph Construction from Medieval Icelandic Sagas: Integrating CIDOC-CRMsoc, Shape Expressions, and Large Language Models

Shintaro Yamada, Ikki Ohmukai
The University of Tokyo, Japan

Keywords: *Knowledge Graph, Shape Expressions, Medieval Icelandic Sagas, Large Language Models, CIDOC-CRM*

Medieval Icelandic sagas provide valuable insights into the society of their time by depicting intricate social and kinship relationships among characters, as well as the political dynamics between factions. Recent efforts have begun developing workflows for representing saga narratives as knowledge graphs (Yamada et al., 2024), demonstrating the feasibility of structured semantic representation. However, these initial approaches relied heavily on manual annotation processes, which were extremely labor-intensive and challenging to scale across entire collections. While the recent emergence of large language models (LLMs) has rendered automated processing approaches viable, fully automated methods frequently fail to capture the nuanced relational complexities inherent in historical narratives.

Despite established digital humanities research involving network and spatial analysis of sagas, no comprehensive knowledge graph specifically designed for these narrative texts currently exists. This paper presents a semi-automated methodology for constructing knowledge graphs from saga narratives by utilizing the CIDOC Conceptual Reference Model for Social Phenomena (CRMsoc). CRMsoc is an extension of CIDOC-CRM, designed specifically for modeling social phenomena, based on an ISO standard with over 15 years of development experience in the cultural heritage field. This ontology proves particularly well-suited for modeling social roles, group membership, and social relationships within sagas.

The key innovation of our approach lies in integrating Shape Expressions (ShEx) schemas as both extraction guides and validation mechanisms. The ShEx specification serves dual purposes: first, it provides a machine-readable template that guides the large language models (LLMs) in identifying entities, events, and relations during the extraction phase. Second, it validates the generated RDF triples against the ontological framework to ensure data quality and semantic consistency. This ontology-driven approach enables more accurate and consistent extraction compared to purely data-driven methods.

The processing workflow consists of five stages: 1) creation of ShEx files based on CRMsoc, 2) LLM-based extraction of entities and events from sagas according to the ShEx schema, 3) mapping to CRMsoc properties, 4) generation of RDF triples, and 5) automated ShEx compliance checking followed by expert verification. We conducted a pilot study focusing on approximately 20 chapters from the Íslendinga saga within the larger Sturlunga saga collection.

The resulting knowledge graph successfully captures narrative events, character interactions, social roles, and temporal sequences as interconnected RDF triples. This work constitutes a contribution to computational literary studies, demonstrating how formal ontology and semantic web technologies can facilitate large-scale analysis of historical narratives. As the first application of this combined methodology—integrating CIDOC-CRM/CRMsoc, ShEx-guided extraction, and LLM capabilities—to medieval narratives, our approach holds potential for scalability across entire saga corpora and could be adapted to other medieval narrative traditions.

References

[1] Cimmino, Andrea, Alba Fernández-Izquierdo, and Raúl García-Castro. 2020. "Astrea: Automatic Generation of SHACL Shapes from Ontologies." In The Semantic Web: 17th International Conference, ESWC 2020, Heraklion, Crete, Greece, May 31–June 4, 2020, Proceedings, edited by Andreas Harth, Sabrina Kirrane, Axel-Cyrille Ngonga Ngomo, Heiko Paulheim, Anisa Rula, Anna Lisa Gentile, Peter Haase, and Michael Cochez, 497–513. Lecture Notes in Computer Science 12123. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-49461-2_29.

[2] Dodds, Leigh, and Ian Davis. 2022. Linked Data Patterns: A Pattern Catalogue for Modelling, Publishing, and Consuming Linked Data. Accessed December 10, 2024. https://patterns.dataincubator.org/.

[3] Fernández-Álvarez, Daniel, José Emilio Labra-Gayo, and Daniel Gayo-Avello. 2021. "Automatic Extraction of Shapes Using sheXer." Knowledge-Based Systems 238: 107975. https://doi.org/10.1016/j.knosys.2021.107975.

[4] Hyvönen, Eero, Petri Leskinen, and Jouni Tuominen. 2023. "A Data-Driven Approach to Create an Ontology of Parliamentary Work." In Proceedings of the International Workshop on Semantic Web and Ontology Design for Cultural Heritage Co-Located with the International Semantic Web Conference 2023 (ISWC 2023), edited by Antonis Bikakis, Roberta Ferrario, Stéphane Jean, Béatrice Markhoff, Alessandro Mosca, and Marianna Nicolosi Asmundo. CEUR Workshop Proceedings 3540. https://ceur-ws.org/Vol-3540/paper6.pdf.

[5] Kawamura, Takahiro, Shusaku Egami, Koutarou Tamura, Yasunori Hokazono, Takanori Ugai, Yusuke Koyanagi, Fumihito Nishino, et al. 2020. "Report on the First Knowledge Graph Reasoning Challenge 2018." In Semantic Technology, edited by Xin Wang, Francesca Alessandra Lisi, Guohui Xiao, and Elena Botoeva, 18–34. Lecture Notes in Computer Science. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-41407-8_2.

[6] Lethbridge, Emily. 2020. "Digital Mapping and the Narrative Stratigraphy of Iceland." In Historical Geography, GIScience and Textual Analysis, edited by Charles Travis, Francis Ludlow, and Ferenc Gyuris, 19–32. Historical Geography and Geosciences. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-37569-0_2.

[7] Lethbridge, Emily. n.d. "dataARC Source Repository." GitHub. Accessed December 10, 2024. https://github.com/castuofa/dataarc-source.

[8] Meghini, Carlo, Valentina Bartalesi, and Daniele Metilli. 2021. "Representing Narratives in Digital Libraries: The Narrative Ontology." Semantic Web 12 (2): 241–64. https://doi.org/10.3233/SW-200421.

[9] Ogawa, Jun, Chikahiko Suzuki, Alex Rui Wang, and Kiyonori Nagasaki. 2023. "Collecting Pieces of Historical Knowledge from Documents: Introduction of HIMIKO (Historical Micro Knowledge and Ontology)." Zenodo. https://doi.org/10.5281/zenodo.8107411.

[10] Yamada, Shintaro, Jun Ogawa, and Ikki Ohmukai. 2024. "Representing the Íslendinga Saga as Knowledge Graphs of Events and Social Relationships: Developing Workflows Based on a Pilot Case." Paper presented at the Digital Humanities in the Nordic and Baltic Countries Conference (DHNB 2024), Reykjavík, Iceland, June 2024. https://www.conftool.org/dhnb2024/index.php/YAMADA-Representing_the_%C3%8Dslendinga_Saga_As_Knowledge_Graphs-217.pdf.

16:15–18:00 *POSTER & DEMO* **[30]**

A Digital Anchor: Cultivating Self-Leadership and Personal Agency in Youth through a Spiritual App

Marcella Zoccoli[1], Klea Ziu[2]

[1] *University of Helsinki, Finland*

[2] *Mohamed bin Zayed University of Artificial Intelligence, Abu Dhabi, UAE*

Keywords: *Self-Leadership, Personal Agency, Experiential Learning, Spiritual App, AI, Human-Centered Technology*

This work highlights an ongoing pilot study in response to recent research displaying a growing need for more studies on developing self-leadership in higher education (Park et al. 2023) that use spiritual applications (Toivonen and Gorichanaz 2026). In the current BANI world: Brittle, Anxious, Nonlinear, and Incomprehensible (Syamsir et al. 2025), five generations live and work together (Bailey and Owens 2020) in a fast-paced acceleration of technology and A.I., creating vibrant, hectic dynamics posing challenges and possibilities for leadership development, requiring awareness and discipline in our lives. Young generations seem to be "*lost in abundance*," showing remarkable signs of anxiety, addiction, alienation, scarcity of awareness, and agency on individual and collective levels (Heidt 2024).

To answer the call, rather than viewing digital apps as passive media, we explored their purposeful adaptation to enact deep, reflective, and human-centered experiential learning (Kolb 1984) and how the vision behind the design might enable a "*digital anchor*" grounding young users in processing personal introspection and transformation needed to enhance reflection, the self-leadership process (Ross 2014), and awareness-based personal agency capacity, as *"the capacity to act from a deeper source of knowing and intention"* (Scharmer and Kaufer 2025) in an era of profound digital and social transition.

The application of a Theory U-informed course design framework (Scharmer, 2016) facilitated students' learning by setting aside distractions and suspending judgment, they acted as active participants, not consumers, allowing themselves to turn inward in the "*presencing*" moment, the capacity *"to operate from the emerging future: sensing, tuning in, and acting from one's highest future potential—bringing to life the future that depends on us to bring it into being."* (Presencing Institute 2025).

Our experiment emerged from comparing the qualitative responses of a questionnaire on contemplative practices and app use conducted in 2020 (Zoccoli 2022) and in 2025 in leadership courses, along with researchers' and students' autoethnographies. The first stage was conducted in Spring and Autumn 2025 in a Finnish higher education context. We proposed, as a pedagogical tool, the free, unintrusive,

non-ritualistic spiritual app Miracle of Mind (2025), developed by Isha Foundation, to 74 international students participating in four leadership courses and one cultural event. The app offers 7-minute guided meditation, daily quotes, audios, short videos, and a generative AI interface to ask questions answered by Sadhguru's insights.

Initial readings of the seventy-four students' leadership journaling/feedback, four volunteer-based reviews, and verbal sharing reveal interesting insights. Their voices narrate that after a few meditation sessions, they experienced a positive shift in mood, the ability to stay present, and an increase in awareness of how challenging it is to be still within us; during the process, the inspiring videos provided valuable wisdom. Since the experiment and analysis are still in progress in Spring 2026, we propose to include a third independent observer to provide investigator triangulation. Ensuring the phenomenon is captured with objectivity. The current insights aim to stimulate discussion and further inquiry within the field regarding the use of spiritual apps as pedagogical tools in learning leadership in higher education.

References

Bailey, Elizabeth, and Christina Owens. 2020. Unlocking the Benefits of the Multigenerational Workplace. Harvard Business Publishing. https://www.harvardbusiness.org/wp-content/uploads/2020/08/Unlocking-the-Benefits-of-Multigenerational-Workforces_Aug-2020.pdf.

Haidt, Jonathan. 2024. The Anxious Generation: How the Great Rewiring of Childhood Is Causing an Epidemic of Mental Illness. New York: Penguin Press.

Isha Foundation. n.d. Miracle of Mind. https://isha.sadhguru.org/uk/en/miracle-of-mind.

Kolb, David A. 1984. Experiential Learning: Experience as the Source of Learning and Development. Englewood Cliffs, NJ: Prentice-Hall.

Park, SY, J. Huberty, J. Yourell, KL McAlister, and CC Beatty. 2023. "A Spiritual Self-Care Mobile App (Skylight) for Mental Health, Sleep, and Spiritual Well-Being Among Generation Z and Young Millennials: Cross-Sectional Survey." JMIR Formative Research 7: e50239. https://doi.org/10.2196/50239.

Presencing Institute. 2025. Presence + Sensing. Presencing.https://www.presencing.org/presencing

Ross, S. 2014. "A Conceptual Model for Understanding the Process of Self-Leadership Development and Action-Steps to Promote Personal Leadership Development." Journal of Management Development 33 (4): 299–323. https://doi.org/10.1108/JMD-11-2012-0147.

Syamsir, S., Saputra, N., and Mulia, R. A. 2025. "Leadership Agility in a VUCA World: A Systematic Review, Conceptual Insights, and Research Directions." Cogent Business & Management 12 (1). https://doi.org/10.1080/23311975.2025.2482022.

Scharmer, Otto C. 2016. Theory U: Leading from the Future as It Emerges. Oakland, CA: Berrett-Koehler Publishers.

Scharmer, Otto C., and Katrin Kaufer. 2025. Presencing: 7 Practices for Transforming Self, Society, and Business. Oakland, CA: Berrett-Koehler Publishers.

Toivonen, H. K., & Gorichanaz, T. (2026). Young adults using apps for alternative spirituality. Psychology of Religion and Spirituality. Advance online publication. https://dx.doi.org/10.1037/rel0000606

Zoccoli, M. 2022. "The Intertwining of Spirituality, Business, and Technology – Conscious Leaders on a Conscious Planet." Poster presented at the conference Mind and Matter 2022. Abstract available on page 81. https://www.helsinki.fi/assets/drupal/s3fs-public/from_d7/305494_mind_and_matter_2022_abstracts.pdf

THURSDAY, 12 MARCH 2026

Session 2A — 10:30–12:30

10:30–10:50 *SHORT PAPER* **[31]**

Imprints from the Women's Prison at Christianshavn 1870-1928

Kamilla Matthiassen
Aalborg University, Denmark

Keywords: *Prison history, female prisoners, digital methods, life trajectories, history of experience*

In 1870, Denmark opened its first and only prison exclusively for women in the heart of Copenhagen – at the busy Christianshavn Square. For nearly sixty years, this institution housed thousands of women from across the Danish Kingdom and its colonies. Despite its size, duration, and national reach, the prison has remained largely absent from dominant narratives of Danish penal history. Existing historiography has focused almost exclusively on male penitentiaries such as the Danish male prisons at Horsens and Vridsløselille, heralding them as symbols of penal modernity. In contrast, the women's prison at Christianshavn which functioned from 1870 to 1928 – along with the lives and experiences of those confined within – has been marginalised.

This is not due to a lack of sources, however. The women's prison generated an extraordinary volume of bureaucratic documentation. Central among these is the *General Protocol*, which is a vast hand-written inmate register that was maintained over decades, offering detailed serial biographies of all the women who were incarcerated from 1873 to 1919. Not only, do these entries record the backgrounds of more than 4,000 women, but they also include their behaviour during incarceration and the institutional assessment the prisoners received upon release. Compared to the inmate registers from the male prisons of the period – which are structured and easier to investigate – Christianshavn's general protocol may have been perceived as too unwieldy and messy to be integrated into the established canon of modern prison historiography. This project therefore places the women's prison as a key site for exploring gendered punishment and the experiences of people often left out in historical narratives.

At present, and in response to this challenge, the project remains in an exploratory phase, as I am six months into my PhD. Consequently, this abstract focuses primarily on methodological considerations rather than empirical results. The project is based on a fully transcribed version of the general protocol's thousands of pages. The transcription has been produced using AI-based Handwritten Text Recognition (HTR) via *Transkribus*. Throughout this work, I have trained a custom text recognition model on 900.000 words, achieving a Character Error Rate (CER) of 4%. Therefore, the model is expected to be applicable to the transcription of additional sources from the prison. For now, the completed transcription of the general protocol has been exported from Transkribus and structured into a dataset in which each incarceration constitutes a separate, structured entry.

Beyond such statistical dataset, a selected part of the textual data will be processed using the *Graph of Roles and Action Model* (GRAM). This approach allows for the extraction of *actions*, *actors*, *objects*, and *places* from the narrative biographies contained in the general protocol. These four elements are subsequently linked by their *relationships*, generating a graph database capable of representing life-course trajectories in both chronological and relational ways.

Additionally, the dataset will not remain isolated. Another key component of the methodology is to *link* the information from the general protocol to other relevant sources – including census data, church books, newspapers, or other materials from the prison such as interrogation transcripts, order books, medical journals, or school protocols. This will be done using a combination of manual cross-referencing and computational searching and structuring through the programming language *R*, including the standardisation of names and birthdates and the construction of relational datasets, thereby facilitating the linkage of individuals across multiple sources and administrative systems.

Taken together, these methods enable the reconstruction of multi-layered life histories and helps trace the broader social and institutional networks that shaped and were shaped by the women who were incarcerated. The resulting data structure allows for both close and distant readings, where qualitative micro-histories of selected women can be embedded within larger-scale quantitative patterns. This dual lens makes it possible to identify recurring themes in pathways to incarceration, strategies of resistance or adaptation inside the prison, and post-release trajectories. It also opens space to examine how institutional power operated across domains – linking welfare, criminal justice, and surveillance – and how women navigated these intersecting systems.

On this basis, this method aims to reveal both the shaping influence of institutional structures and the ways women exercised agency within them. Many women entered prison after repeated interactions with orphanages, poor relief systems, workhouses, hospitals, or mental institutions – a pattern frequently recorded in the general protocol. The inmate's biographies are therefore not simply accounts of crime and punishment but also document ongoing negotiations with multiple coercive institutions. Yet even inside prison, where they were disciplined, classified and assessed, the women formed relationships, exerted agency, and developed strategies for survival. This focus thus contributes to the *history of experience* by centring the embodied, affective, and contingent dimensions of punishment. Rather than viewing the prison at Christianshavn as a footnote to the male penitentiary story, it is presented as a core part of Denmark's carceral history – one that reveals the gendered textures of penal modernity and its historical patterns. Hence, the project argues for a thorough and critical reading of the records – not as passive bureaucratic residues, but as layered and conflicted texts through which incarcerated women can be heard.

References

Annola, Johanna & Lindberg, Hanna & Markkola, Pirjo (ed.): "Lived Institutions as History of Experience". Palgrave Studies in the History of Experience (2024)

Bak, Greg & Rostgaard, Marianne (red): The Nordic Model of Digital Archiving. Routlegde, 1st edition (2023), chapter 9: Mathiesen, Nicolai Rask & Revuelta-Eugercios, Barbara & Robinson, Olivia & Thomsen, Asbjørn Romvig: "Transforming archival records into historical big data. Visualising human and computer processes in the Link-Lives project" (p. 152-172)

Chevaleyre, Claude & Heinsen, Johan & Schiel, Juliane & Peres, Corinna: "Graph of Roles and Actions Model". Online publication (2024). URL: https://github.com/cchevale/GRAM_public/tree/main (visited 18-12-2025)

Foucault, Michel: Overvågning og straf. Fængslets fødsel. DET lille FORLAG, 2002 (1975)

Heinsen, Johan: "Carceral chains: pathways through a convict labour institution, 1690-1830". Scandinavian Journal of History (2023)

Ignatieff, Michael: A Just Measure of Pain: The Penitentiary in the Industrial Revolution (1978)

Johnston, Helen: "Imprisoned mothers in Victorian England, 1853-1900: Motherhood, identify and the convict prison". University of Hull (2018)

Menis, Susanna: A History of Women's Prisons in England: The Myth of Prisoner Reformation. Cambridge Scholar Publishing (2020)

Rasmussen, Leonora Lottrup: "Kærlighed i Fængsel. Et studie af intime relationer i Horsens Straffeanstalt i slutningen af 1800-tallet". Kulturstudier, Nr. 2 (2020)

Rigsarkivet, Arkivalieronline (RA): Statsfængslet i Horsens: Stambog 1853-1909; RA: Statsfængslet i Vridsløselille: Stamrulle 1859-1932

Smith, Peter Scharff: Moralske Hospitaler – det moderne fængselsvæsens gennembrud 1770-1870. Forum (2003)

Stuckenberg, Frederik: Fængselsvæsenet i Danmark 1742-1839. En Historisk Skildring. København. I Kommission hos Universitetsboghandler G.E.C. Gad, bind 2 (1896)

Valentin, Emilie Luther: Feelings of Imprisonment – Experiences from the prison workhouse at Christianshavn, 1769-1800. Aalborg Universitet (2022)

Zedner, Lucier: "Women, Crime and Penal Responses: A Historical Account". The University of Chicago (1991)

10:50–11:20 *LONG PAPER* **[32]**

Linking Heterogeneous Historical Sources: A Machine-Learning Approach to Danish Census and Population Data (1880-1921)

Tobias Kallehauge[1], Asbjørn Romvig Thomsen[1], Olivia Robinson[1], Anne Løkke[2], Barbara A. Revuelta-Eugercios[1]

[1] *Danish National Archives (Rigsarkivet), Denmark*

[2] *University of Copenhagen, Denmark*

Keywords: *Historical record linkage; machine-learning; Denmark; census records*

1. Introduction

Over the past decade, the digital transformation of archives has generated unprecedented research opportunities. Millions of historical records have undergone transcription — through new handwritten text recognition (HTR) technologies (Terras 2022), numerous ad hoc digitization projects (Colavizza et al. 2021), and the collective work of volunteers in citizen-science and crowdsourcing initiatives (Prats López et al. 2025). This abundance of digital material has enabled the creation of historical databases containing information on individuals across a vast array of heterogeneous sources. Examples in the last decade include large-scale projects in Sweden (*SwedPop – A national research infrastructure* 2023), Norway (Thorvaldsen, Andersen, and Sommerseth 2015), Scotland (Scottish Centre for Administrative Data Research (SCADR), n.d.), the US (Helgertz et al. 2022), France (Boillet et al. 2024), and Denmark (Revuelta Eugercios, Robinson, and Løkke 2021). One of the key technologies behind these databases is *automated record linkage*, i.e., computer-based methods for connecting person-level information scattered across different documents and time periods. As opposed to manual record linkage, automated methods can be scaled up to link an entire population with millions of records, and over the past twenty years, they have evolved from rule-based matching to sophisticated computational approaches based on machine-learning (Ruggles, Fitch, and Roberts 2018).

These advances have opened up new and exciting applications, but they have also amplified existing challenges in record linkage. First, the automated methods must still be carefully tuned for linking two data sources (e.g., parish records to a census), and it is seen that the heterogeneity of sources has increased dramatically as more sources have been transcribed (e.g., newspapers and memoirs). Second, using domain expertise with knowledge of the source and historical period is becoming more important throughout the entire pipeline of automated linking, not only for making training data with examples of links and non-links, but also for designing the model, for assessing if the patterns learned by the model for linking align with the historical context, and for validating the resulting links.

This article examines the case of linking individuals across the Danish censuses 1880-1921 and the Copenhagen Police Sheet Register (CPS) (Politiets Registerblade, 1890-1923), which all differ both in terms of format and methods of transcription. The central research question is: How can a machine-learning approach informed by domain expertise link heterogeneous historical datasets at scale? To address this question, the article presents the workflow developed within the Link-Lives project (Revuelta Eugercios, Robinson, and Løkke 2021). The following section reviews the state of the art in historical record linkage, section 3 introduces the Danish dataset, section 4 details the machine-learning approach, and section 5 presents the main results. Finally, section 6 concludes the paper.

2. State of the Art

The field of automated record linkage emerged with the first computer-readable historical datasets, when scholars began exploring how digital methods could connect individual records across sources (Ruggles, Fitch, and Roberts 2018). One of the earlier and well-cited works in this field is (Ruggles 2002), which uses a set of *hand-crafted* rules (i.e., rule-based) to determine links with a careful choice of variables to avoid unrepresentative linking. Another approach is *supervised machine learning*, where

training data is used to construct a model (often probabilistic, such as probit) that mimics the examples seen during training (Minardi, Greco, Barban, et al. 2025).

All through the 2000s and 2010s, rule-based and probabilistic linkage frameworks were prevalent. Many projects have employed purely automatic methods (Mandemakers et al. 2023), but many adopt semi-automatic methods in which a large part of the record linkage was automatic, while leaving the uncertain cases for human review (Vézina and Bournival 2020; Thorvaldsen, Andersen, and Sommerseth 2015; Edvinsson, Westberg, and Engberg 2016). In the last decade, more advanced supervised methods have appeared, including support vector machines (Fu et al. 2014; Antonie et al. 2020), gradient boosting (Park 2022), and knowledge graphs (Gautam et al. 2020). These techniques can model non-linear relationships among variables and achieve higher performance, but they require high-quality training data and careful calibration to perform well. Another challenge is linking the increasingly large datasets created by HTR, where scaling a model to millions of records can be difficult. The use of HTR also increases risks of omissions and suppression of low-frequency values (Nockels, Gooding, and Terras 2024), which need to be accounted for when designing a model. At the time of writing, there are no published examples of large-scale record linkage involving HTR-generated datasets.

3. Data: Linking Heterogeneous Sources

3.1. Datasets to be linked

The study uses Danish censuses (1880–1921) and the CPS (1890–1923). These datasets differ in origin, content, and transcription method. The censuses were all nationwide events, each carried out on a single day, and collected information for entire households. The CPS, on the other hand, was only for Copenhagen and dynamically tracked the city's population between 1890 and 1923 on sheets, tied to a main person with husband/wife and children, also registered on the sheet (Københavns Stadsarkiv). The information in the sources is similar, with name, birth date/age, and birthplace among the key variables available in both the censuses and the CPS (with some exceptions). The sources were transcribed as follows:

- Censuses 1880 and 1901: transcribed manually by volunteers as part of the National Archives Crowdsourcing project, Danish Demographic Database (Clausen 2015).

Table 1
Main characteristics of the censuses to be linked.

Dataset	Transcription type	Number of records	Key variables	Coverage
Census 1880	Crowdsourced	1,979,455	Name, age, birthplace	National
Census 1890	HTR	2,187,404	Name, age, birthplace	National
Census 1901	Crowdsourced	2,468,040	Name, birth date, birthplace	National
Census 1911	HTR	2,742,841	Name, birth date, birthplace	National
Census 1916	HTR	2,767,349	Name, birth date (no birthplace)	National
Census 1921	HTR	3,318,554	Name, birth date, birthplace	National
CPS	Crowdsourced	1,965,256	Name, birth date, birthplace, address	Copenhagen

- Censuses 1890, 1911, 1916, and 1921: automatically transcribed with HTR by Rooftop as part of the Link-Lives project in the period 2022–2023.
- Copenhagen Police Sheet Register: transcribed manually by volunteers crowdsourced at the Copenhagen City Archives (Københavns Stadsarkiv).

The main details of the sources are summarized in table 1.

3.2. Benchmark Data for Training, testing and validation

To create training data, the project manually linked samples between the different datasets. Two independent linkers were given the same randomised sample of records in one source to be linked to another (e.g., 1890 to 1880) using a custom interface that automatically suggests potential links, but also allows for searching the entire dataset. Then, disagreements were resolved by a third person (the arbiter) following the approach described in (Revuelta-Eugercios, Robinson, Mathiesen, et al. 2025).

The domain experts carrying out the manual linking were highly trained historians who have been engaged in the project over a long period.

This approach was used to generate 2,807 manual link decisions between censuses (1921 → 1916 → 1911 → 1901 → 1890 → 1880) and 2,004 manual link decisions for CPS to censuses (CPS → 1921, 1916, 1911, 1901). The proportion of manual link decisions requiring arbitration was 3.6% and 4.9% for census-to-census and CPS-to-census, respectively. In addition, after the automatic linking was complete, a post-validation effort was carried out by sampling at least 100 automatically generated links for each of the source pairings (in total 1,300), which were then verified using the same interface.

4. Methods

4.1. Modelling Approach

The general workflow consisted of harmonization to improve data comparability, blocking to reduce the number of potential links, feature encoding, model fitting on encoded training data, full-scale prediction between all source pairs using the trained model, calibration of link thresholds, and evaluation in terms of precision, recall and link rate[5].

XGBoost was used as the model for link prediction (Chen 2016), with three separate trained models for the nine source pair combinations:

- **C1921–1901** model linking censuses 1921 → 1916 → 1911 → 1901 (with birth date)
- **C1901–1880** model linking censuses 1901 → 1890 → 1880 (only age)
- **CPS–C** model linking CPS to censuses 1921, 1916, 1911, 1901

Table 2

Aggregate performance of models. 95% confidence intervals (Wilsons) are shown for precision and recall.

Model	Source pairs	Precision [%]	Recall [%]	Link rate [%]
C1921-1901	1921 → 1916 → 1911 → 1901	97.5 (95.9, 98.5)	50.7 (45.5, 56.9)	40.7
C1901-1880	1901 → 1890 → 1880	98.5 (95.7, 99.5)	41.4 (36.1, 46.8)	34.1
CPS-Census	CPS → 1901, 1911, 1916, 1921	98.8 (97.4, 99.4)	-	62.6

4.2. Blocking and sampling non-links

Blocking was used to reduce the search space by restricting candidate pairs to the same sex and an absolute birth-year difference ≤ 5 for C1901-1880 and Levenshtein distance ≤ 1 for C1921-1901. After blocking, the amount of training data (which mostly contained positive links) was increased by sampling multiple non-links using the fact that a positive link to a census rules out any other positive links, which are then necessarily non-links. It was found that the addition of non-links helps the model to better mimic the linking approach of the manual domain-expert linkers, e.g., learning not to link if names are too dissimilar, which ultimately gives higher precision in the predicted links.

4.3. Future encoding

One of the key steps in the workflow is designing a feature encoding function where tabular data from pairs of records in two sources are encoded as numerical vectors characterizing the similarity in key variables (name, birth date, etc.). Encoding is done primarily for variables on an individual level, but variables of other individuals in the household are also given as support information (e.g., a child with

[5] Precison = TP/(TP + FP), Recall = TP/(TP + FN), where TP is true positives, FP is false positives, and FN is false negatives. Link rate = “number of links”/“number of linkables” where the number of linkables counts individuals that, according to birth date/age, were born before the year of the source they were to be linked to.

variables from the first male and female in the household as support). Several methods are used for encoding information, including Jaro–Winkler similarity for comparing string similarity, absolute numerical distance between numbers, and Levenshtein distance. It was found that the latter was particularly useful for comparing years or dates transcribed by HTR since it counts the number of "edits" between two strings, which better captures transcription errors than numerical distance (e.g., 1901 being transcribed as 1801). Following these and other methods, the similarity of key variables between sources listed in table 1 was encoded as numbers. A data-driven approach was used to identify significant features by first suggesting a large number of features and then using sequential feature floating selection (SFFS) to reduce the number. It was found that the XGBoost models performed well with a large number of (potentially redundant) features, due to a richer representation of potential links, which can help mitigate errors in the data, e.g., introduced by HTR. The final models used 49 features for C1921–1901, 57 for C1901–1880, and 34 for CPS–C.

4.4. Training, Thresholds, and Evaluation

The XGBoost models were fitted based on the encoded training data, where weights were applied to correct for class imbalance between links and non-links. The trained models were then used to predict link scores between 0 and 1 between each source pair. Finally, a series of rules was used to resolve links by tuning different thresholds on the scores to achieve the desired precision (at the cost of recall).

5. Results

The models achieved a precision of around 97.5% or higher, recall between 41.4% and 50.7%, and link rates between 40.7 and 62.6% as seen in table 2. Note that precision is calculated based on the aforementioned post-validation effort, while recall is based on test data (missing for CPS-Census).

The performance for specific source pairs varied, likely due to missing variables in some sources and the varying quality of the HTR transcriptions. The performance for source pairs involving the 1916 census was the lowest, due to the 1916 census missing birthplace and having a lower quality transcription. Source pairs involving the 1901 census tended to perform well, likely due to having both birth dates and places, and being manually transcribed. The CPS–Census model performed well in terms of precision. The link rate from CPS to any given census was between 16.8% and 41.1%, but with 62.6% of linkable records in CPS being linked to at least one census.

References

Antonie, Luiza, Kris Inwood, Chris Minns, and Fraser Summerfield. 2020. "Selection Bias Encountered in the Systematic Linking of Historical Census Records." Social Science History 44 (3): 555–70. https://doi.org/10.1017/ssh.2020.15.

Boillet, Mélodie, Solène Tarride, Yoann Schneider, Bastien Abadie, Lionel Kesztenbaum, and Christopher Kermorvant. 2024. "The Socface Project: Large-Scale Collection, Processing, and Analysis of a Century of French Censuses." In Document Analysis and Recognition - ICDAR 2024, edited by Elisa H. Barney Smith, Marcus Liwicki, and Liangrui Peng. Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-70543-4_4.

Clausen, Nanna Floor. 2015. "The Danish Demographic Database—Principles and Methods for Cleaning and Standardisation of Data." In Population Reconstruction, edited by Gerrit Bloothooft, Peter Christen, Kees Mandemakers, and Marijn Schraagen. Springer.

Colavizza, Giovanni, Tobias Blanke, Charles Jeurgens, and Julia Noordegraaf. 2021. "Archives and AI: An Overview of Current Debates and Future Perspectives." J. Comput. Cult. Herit. 15 (1): 4:1-4:15. https://doi.org/10.1145/3479010.

Edvinsson, Sören, Annika Westberg, and Elisabeth Engberg. 2016. "A Unique Source for Innovative Longitudinal Research: The POPLINK Database." Historical Life Course Studies 3 (March): 20–31.

Fu, Zhichun, H.M. Boot, Peter Christen, and Jun Zhou. 2014. "Automatic Record Linkage of Individuals and Households in Historical Census Data." International Journal of Humanities and Arts Computing 8 (2): 204–25. https://doi.org/10.3366/ijhac.2014.0130.

Gautam, Bhaskar, Oriol Ramos Terrades, Joana Maria Pujadas-Mora, and Miquel Valls. 2020. "Knowledge Graph Based Methods for Record Linkage." Pattern Recognition Letters 136 (August): 127–33. https://doi.org/10.1016/j.patrec.2020.05.025.

Helgertz, Jonas, Joseph Price, Jacob Wellington, Kelly J. Thompson, Steven Ruggles, and Catherine A. Fitch. 2022. "A New Strategy for Linking U.S. Historical Censuses: A Case Study for the IPUMS Multigenerational Longitudinal Panel." Historical Methods 55 (1): 12–29. https://doi.org/10.1080/01615440.2021.1985027.

Mandemakers, Kees, Gerit Bloothooft, Fons Laan, Joe Raad, Rick J. Mourits, and Richard L. Zijdeman. 2023. “LINKS. A System for Historical Family Reconstruction in the Netherlands. H.” Historical Life Course Studies 13: 148–85. https://doi.org/10.51964/hlcs14685.

Nockels, Joseph, Paul Gooding, and Melissa Terras. 2024. “The Implications of Handwritten Text Recognition for Accessing the Past at Scale.” Journal of Documentation 80 (7): 148–67. world. https://doi.org/10.1108/JD-09-2023-0183.

Park, Narae. 2022. “Record Linkage of Norwegian Historical Census Data Using Machine Learning.” Master Thesis, The Artic University of Norway. https://munin.uit.no/handle/10037/28399.

“Politiets registerblade.” n.d. Københavns Stadsarkiv. Accessed March 25, 2021. https://kbharkiv.dk/brug-samlingerne/kilder-paa-nettet/politiets-registerblade/.

Prats López, M., Van Oort ,T., Ganzevoort ,W., Van Galen ,C., and R. J. and Mourits. 2025. “Understanding Patterns of Engagement in the Citizen Humanities: The Civil Records of Suriname.” Historical Methods: A Journal of Quantitative and Interdisciplinary History 58 (1): 1–16. https://doi.org/10.1080/01615440.2024.2414925.

Revuelta Eugercios, Revuelta Eugercios, Bárbara Ana, Olivia Robinson, and Anne Løkke. 2021. “Link-Lives, Historical Big Data: Reconstructing Millions of Life Courses from Archival Records Using Domain Experts and Machine Learning.” Proceedings of Linked Archives International Workshop 2021 Co-Located with 25th International Conference on Theory and Practice of Digital Libraries (TPDL 2021) 3019: 135–43. http://ceur-ws.org/Vol-3019/LinkedArchives_2021_paper_9.pdf.

Revuelta-Eugercios, Bárbara, Olivia Robinson, Nicolai Mathiesen, et al. 2025. Link-Lives v.2 Guide.

Ruggles, Steven. 2002. “Linking Historial Censuses: A New Approach.” History & Computing 14 (1/2): 213–24. a9h (22850908).

Ruggles, Steven, Catherine A. Fitch, and Evan Roberts. 2018. “Historical Census Record Linkage.” In Annual Review of Sociology, vol. 44. no. Volume 44, 2018. Annual Reviews. https://doi.org/10.1146/annurev-soc-073117-041447.

“Scottish Historic Population Platform (SHiPP) | SCADR.” n.d. Accessed October 30, 2024. https://www.scadr.ac.uk/our-research/shipp.

“SwedPop – A national research infrastructure.” 2023. January 12. https://swedpop.se/.

Terras, Melissa. 2022. “Chapter 7: Inviting AI into the Archives: The Reception of Handwritten Recognition Technology into Historical Manuscript Transcription.” In Archives, Access and Artificial Intelligence. Bielefeld University Press. https://www.degruyterbrill.com/document/doi/10.1515/9783839455845-008/html?lang=en.

Thorvaldsen, Gunnar, Trygve Andersen, and Hilde L. Sommerseth. 2015. “Record Linkage in the Historical Population Register for Norway.” In Population Reconstruction, edited by Gerrit Bloothooft, Peter Christen, Kees Mandemakers, and Marijn Schraagen. Springer.

Vézina, Hélène, and Jean-Sébastien Bournival. 2020. “An Overview of the BALSAC Population Database. Past Developments, Current State and Future Prospects.” Historical Life Course Studies, ahead of print, August 25. https://doi.org/10.51964/hlcs9299.

11:20–11:40 *SHORT PAPER* **[33]**

Migrate and visualise dispersed digitized data: A case study of seafarer’s voyages and Estonian seafarers on Norwegian ships during the Second World War

Ole Jørgen Søndbø Abrahamsen

Centre for the History of Seafarers at War, ARKIVET Peace and Human Rights Centre, Kristiansnd, Norway

Keywords: *Krigsseilerregisteret, Digital history, Database*

This paper examines and compares different methods for collecting and presenting existing dispersed, digitized data concerning the seafarers on Norwegian ships during the Second World War in the online database *Krigsseilerregisteret.no*. To do this, I will examine two cases from my work with *Krigsseilerregisteret* and discuss the impact of the different methods on the structure and visualisation of the data in the *Krigsseilerregisteret* after it has been collected and imported into the database. I will argue that the methods, based on different technology and collaboration, are not neutral processes without implications. Hence, the historian should be aware that both the technology and persons involved in the development of the database will influence the result. By comparing the two specific cases, the paper`s analysis of the relationship between method, collecting dispersed data and visualisation offer valuable insights for digital history projects across the Nordic and Baltic regions.

The two cases are from my work with the online database *Krigsseilerregisteret*. This database has for ten years been developed by historians at the Centre for History of Seafarers at War, at ARKIVET Peace and Human Rights Centre in Kristiansand, Norway. The initial method that was used to build the database was to engage volunteers to transcribe information from digitized archive sources into the database and then use historians to quality assure the information. The goal of the database is to register every seafarer on Norwegian ships, and Norwegian seafarers on foreign ships during the Second World War. The database is a history from below, where every individual seafarer gets their own profile with information about their wartime effort. The context discussed in the paper is relevant for historical databases on individuals as such, and especially for the increasing number of databases on individuals from the Second World War.

The first case covers the experiences from migrating and registering data about where more than 1000 Norwegian merchant ships sailed during the war for the Allies. When combining this information with the existing information in *Krigsseilerregisteret* on when the 60,000 seafarers were serving on different ships, it was possible to visualize on maps – both on a micro and macro level – where the seafarers sailed during the Second World War. To achieve this, we employed a combination of methods: web scraping, AI transcription, and traditional transcription by volunteers. The paper will discuss the advantages and disadvantages of the methods used and how they affected the visualisation.

The second case is based on a transnational collaboration project between *Krigsseilerregisteret* and volunteers in Estonia, who have searched for archival sources on Estonian seafarers on Norwegian ships in the National Archive of Estonia. This is done to supplement, and quality assure the existing information in the database, which is primarily based on Norwegian archival sources. This case provides a clear example of how collaboration and a new archival context can alter the materiality of the source, adding transnational layers that were absent from the Norwegian records alone.

In the discussion of what happens to a historical source when it is digitised and processed with the help of computing, Walter Benjamin's *The Work of Art in the Age of Mechanical Reproduction* is relevant. In his essay, Benjamin argues that works of art take on a whole new materiality the moment the possibility of reproduction arises (Benjamin 1975). This argumentation and analysis are also useful starting points for historians who seek to assess how the sources are affected by digitization and digital processes which in a sense try to reproduce the original source (Bastiansen 2023). Depending on the type of method you are using to create a database, the materiality of the sources will change. The cases in this paper are therefore relevant for a broader discussion about how digital tools shape our study of the past. Both cases look at projects where digitized source-material is transformed, using different methods, to correspond in the same database Krigsseilerregisteret. Together, the two cases illustrate that different methods of migrating dispersed data leave different material traces in the database – traces that are not merely technical but makes material significant conditions for the digital visualization of the data.

Applying Benjamin's framework, the paper will examine how the materiality of a digitized archive source changes when it is: 1) manually transcribed by a human (introducing human interpretation); 2) transcribed by AI (introducing computing and automated processes); 3) a result of web scraping (introducing the collaboration with IT engineers); or 4) cross-referenced by a foreign volunteer in a different archival context (gaining new, transnational layers). I will argue, based on the two concrete cases, that these shifts in materiality before the information is migrated as data into the same database directly enable or constrain the possibilities when it comes to visualisations of the data after the migration. I will also discuss how particularly important elements can be “lost in translation” both directly and indirectly.

To underline the papers significance, I will build on Adam Crumble’s book *Technology and the Historian*. He argues that creating an academic vocabulary that can help define what "Digital History" is, will best be developed by first examining the history of the field in its context. His main argument is that there is not one story about technology and the historian, but many (Crymble 2021). As a result, it is therefore important that technology’s impact on our work as historians is examined concretely in its relevant context. Bastiansen has pointed out that this operation is scarcely touched upon in a Norwegian context (Bastiansen 2023). By examining these specific cases, the paper aims to contribute to the discussion about technology’s impact on databases containing information from digitized sources from the past in the Nordics and the Baltics, as called for by both Crymble and Bastiansen.

References

Bastiansen, Henrik G. 2023. Når fortiden blir digital: Medier, kilder og historie i digitaliseringens tid. 1st. Oslo. Universitetsforlaget.

Benjamin, Walter. 1975. Kunstverket i reproduksjonsalderen og andre essays. 1st. Oslo. Gyldendal Norsk forlag.

Crymble, Adam. 2022. Technology and the Historian: Transformations in the Digital Age, 1st. Illionois. University of Illinois press.

11:40–12:00 *SHORT PAPER* **[34]**

From Transcription to Meaning: Digital Methods for Unlocking Early Modern Danish Court Records

Jeppe Büchert Netterstrøm, Kristian Pindstrup

Aarhus University

Keywords: *AI-based transcription, early modern Denmark, historical records, digital text analysis, language modeling*

In recent years, large-scale digitisation and AI-based transcription have begun to transform early modern research. Thousands of pages of handwritten historical records, once inaccessible due to paleographic barriers, are now becoming machine-readable text through tools such as Transkribus (e.g. Heinsen & Bøgeskov 2025). This paper presents methodological insights from two projects at Aarhus University: *Using artificial intelligence to challenge state evolutionism: Homicide rates and patterns in 16th- and 17th-Century Denmark* (Aarhus University Research Foundation, 2024) and *Unlocking the 16th Century: AI-based Transcription and Online Publication of Early Modern Danish Law Court Records* (Carlsberg Foundation, 2025–26).

The early modern Danish court material is among the richest legal archives in Northern Europe, yet its scope, tens of thousands of manuscript pages, has long made systematic study difficult and time-consuming. The use of Transkribus has made it feasible to transcribe this material at scale. The *Using artificial intelligence* project (2024) focuses on methodological innovation: How machine learning can be used to revisit long-standing narratives about violence, conflict, and legal culture, based on a relatively narrow source material (homicide cases from the high court of Jutland 1620-1660). The Carlsberg-funded continuation (2025-26) expands this approach, and projects it back in time, through large-scale digitisation of sixteenth-century records from multiple jurisdictions, providing both a foundation for new historical analysis and a resource for future digital-humanities research. Both projects use Transkribus as transcription tool. New Transkribus models have been trained to handle sixteenth- and seventeenth-century Danish handwriting (accuracy 90+ percent), and about 13,000 pages of seventeenth-century high-court material from Viborg and 45,000 pages of sixteenth-century court records from all over Denmark have been transcribed (Netterstrøm & Pindstrup 2024a; 2024b; 2024c; 2025; 2026)

AI-based transcription changes not only access but also method. Once text is machine-readable, researchers can move beyond close reading and word search to pattern recognition across millions of words: Identifying shifts in legal vocabulary, procedural change, or evolving social categories such as gender and status. A next methodological step is the construction of a language model trained specifically on early modern Danish. Such a model can capture historical semantics and spelling variation, enabling more nuanced search and clustering. The paper will briefly discuss ongoing preparatory work toward this goal (Enevoldsen et al. 2021; Manjavacas & Fonteyn 2021, 2022; Al-Laith et al. 2024).

The transcribed corpora support both traditional and computational analysis. They enable qualitative as well as quantitative studies of a range of historical phenomena. The integration of handwritten-text recognition and linguistic modeling provides a case study in how humanities scholars can build infrastructure while pursuing substantive research questions.

The paper situates the Danish projects within a broader European movement to digitise and computationally analyse premodern records (e.g. Heinsen & Bøgeskov 2025). Yet it argues that such work must remain source-critical: Transcription errors, variant spellings, and local idioms require historians' domain expertise. The combination of archival, historical, linguistic, and technical knowledge, rather than automation alone, determines the analytical quality of AI-derived corpora. The initiatives presented in this paper demonstrate how digital methods can open an entire historical domain that was

previously closed to large-scale systematic inquiry. By pairing transcription with emerging language-model technology, they illustrate a pragmatic, historically informed approach to AI in the humanities: Not replacing interpretation, but scaling and sharpening it.

References

Al-Laith, A., Conroy, A., Bjerring-Hansen, J. & Hershcovich, D. 2024. Development and Evaluation of Pretrained Language Models for Historical Danish and Norwegian Literary Texts. https://aclanthology.org/2024.lrec-main.431

Enevoldsen, Kasper, Hansen, Lasse & Nielbo, Kristoffer L. 2021. DaCy: A Unified Framework for Danish NLP. Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa 2021). https://arxiv.org/abs/2107.05295.

Heinsen, Johan & Bøgeskov, Camilla. 2025. A World in Print: Introducing a Danish-Norwegian corpus of historical newspapers. https://arxiv.org/abs/2509.02356

Manjavacas, Enrique & Fonteyn, Lauren. 2021. GysBERT: a Dutch BERT Model Trained on Historical Corpora. Presented at NLP4DH Workshop 2021. https://macberth.netlify.app. https://huggingface.co/emanjavacas/GysBERT.

Manjavacas, Enrique & Fonteyn, Lauren. 2022. Adapting vs. Pre-training Language Models for Historical Languages. Journal of Data Mining & Digital Humanities. https://jdmdh.episciences.org/9690.

Netterstrøm, Jeppe Büchert & Pindstrup, Kristian. 2024a. Danish 17th Century 4.0: Transkribus model. Aarhus University.

Netterstrøm, Jeppe Büchert & Pindstrup, Kristian. 2024b. Drabssager fra Viborg landstings dombøger 1622-1666: Transkribus collection. Aarhus University.

Netterstrøm, Jeppe Büchert & Pindstrup, Kristian. 2024c. Viborg landstings tingbøger 1619-1666: Transkribus collection. Aarhus University.

Netterstrøm, Jeppe Büchert & Pindstrup, Kristian. 2025. 16th Century Danish Court Records. Transkribus model. Aarhus University.

Netterstrøm, Jeppe Büchert & Pindstrup, Kristian. 2026. Danish 16th and 17th Century Administrative Handwriting: Transkribus model. Aarhus University

12:00–12:30 *LONG PAPER* **[35]**

Signals from the Field: A Study of Digital Practices and Needs in Sweden

Julia Kuhlin, Daniel Ihrmark, Koraljka Golub, Ahmad Kamal

Linnaeus University, Sweden

Keywords: *Huminfra, Research infrastructures, Digital practice, Digital competence, Training needs*

Introduction and Background

The Swedish national research infrastructure Huminfra is a collaboration among twelve partner universities and cultural heritage institutions, dedicated to supporting and developing digital research in the humanities and social sciences. As part of its continued mission for 2025–2028, Huminfra has conducted a nationwide survey to gather empirical insights into researchers' digital practices, needs, and challenges. In parallel, Sweden's membership in DARIAH-EU entails a responsibility to contribute to and align with broader European efforts to understand and strengthen digital scholarly practices

The study seeks to answer three interrelated questions:

1. How are digital tools and resources used by researchers and cultural heritage professionals in Sweden, and what levels of digital competence characterize their work?
2. What training needs are expressed with respect to digital methods or resources?
3. How do researchers and professionals discover and access digital tools, resources, and training opportunities, and what is their level of awareness of infrastructures such as Huminfra and DARIAH-SE?

The study represents the first systematic attempt to chart the digital landscape of humanities and social science research, as well as cultural heritage work, in Sweden since the field's expansion over the past decade/s. By aligning its design with the survey conducted by Digital Methods and Practices Observatory (DiMPO) working group of DARIAH-EU (Dallas & Chatzidiakou, 2022), this study ensures

comparability with similar national initiatives in Austria, Finland, Greece, Lithuania, Poland, Serbia, and Switzerland.

The DiMPO survey (Dallas & Chatzidiakou, 2022), with over 2,000 respondents from seven European countries, provides a key reference for situating Sweden within the broader digital humanities landscape. It showed that digital engagement is now widespread: 87.1% of researchers use digital tools to collect or create research assets, 73.3% to process or analyze them, and 65.5% to curate or enrich data (p. 34). Despite this broad uptake, respondents identified major needs for improved access to existing digital resources and the digitization of non-digital materials (pp. 48–49). These insights underscore the relevance of the Swedish survey, which aims to capture how such patterns manifest within the Swedish national research environment and how Huminfra and DARIAH-SE might address them.

Methodology

Survey Design

The Swedish survey largely follows the structure of the DiMPO questionnaire, addressing five areas: (1) Respondent background, (2) Current use of digital tools, (3) Current use of digital resources (4) Training needs (5) Awareness and discovery practices. The survey combines closed and open-ended questions to balance comparability and depth.

Sampling and Data Collection

The survey was made available online in May 2025 and closed on September 30, 2025. The sample included both researchers directly affiliated with Huminfra nodes and those working in universities, archives, museums, and libraries outside the formal consortium.

In total, 209 responses were collected: 196 representing 26 higher education institutions across Sweden, and 13 from a range of cultural heritage institutions, including museums, archives, and libraries. Among respondents working in higher education institutions, the primary field of research was reported as follows: 96 in the humanities and the arts, 75 in the social sciences, 3 in the natural sciences (computer science), and 1 in the medical and health sciences. Twenty-one respondents reported fields extending beyond one field of research, namely 15 in both the humanities and social sciences, and 6 in the humanities and natural sciences (all in computer science).

Data Analysis

At this stage, only the quantitative data from closed questions have been analyzed. Hence, the current analysis focuses on identifying broad patterns in tool and resource use, self-assessed competence, training demand, and awareness of infrastructures. Before the conference, cross-tabulations and qualitative thematic coding will follow.

Results (due to word limitations, only part of the results are reported here)

Use of Digital Tools

Among the 209 respondents, 161 reported that they used digital tools beyond basic applications such as email, web browsing, or word processing in their work. The most frequently mentioned tools used in work were Excel and NVivo, followed by large language models (LLMs), R, Python, ArcGIS, Zotero, and Gephi. Self-assessed proficiency levels varied widely: advanced (23), very good (23), adequate (83), modest (58), and insufficient (22). In total, roughly 38% considered their competence below adequate, revealing a pronounced digital skills divide.

As for the challenges reported, these included steep learning curves and the time required to become proficient, software limitations (particularly in qualitative data analysis tools such as NVivo and Atlas.ti) linguistic constraints, (as many text analysis tools are optimized for English rather than Swedish), and difficulties in balancing functionality and usability, with some tools perceived as too complex for occasional users.

Use of Digital Resources

A total of 158 respondents reported working with digital resources. Awareness levels were self-rated as comprehensive (12), very good (61), adequate (66), modest (51), or insufficient (19), hence around one-third thus assessed their awareness as below sufficient.

The types of data most frequently used were textual (143), followed by images (76), metadata (52), numerical data (47), video (39), geospatial (24), and audio (8). This demonstrates a continuing textual orientation within Swedish humanities research, although multimodal and spatial materials are gaining ground.

Respondents primarily used digital resources for analyzing (137), collecting (109), and accessing (100) data, followed by processing or digitizing (90) and managing (51). Activities such as enriching (41), maintaining (32), or publishing datasets were far less common, suggesting that Swedish researchers act mainly as users of existing data rather than producers of new digital corpora.

Training Needs

One of the most striking finding concerns training. Among those using digital tools, two-thirds (105 of 161) had received no formal instruction, and only 58 of all 209 respondents had ever received training in working with digital resources. Consequently, much digital competence is acquired informally, through self-learning or peer exchange.

When asked about desired areas of training, the strongest interest was expressed in: Analysis, organization, and visualization of data (147), digitization of non-digital materials (142), and data collection and processing from digital sources (122 respondents)

Preferences regarding training formats varied, but the most popular options were on-site workshops (125) and online workshops (96), followed by recorded lectures (82), walkthroughs (78), live lectures (70), and Massive Open Online Courses (MOOCs) (48). These responses indicate a preference for interactive and practice-oriented formats, although lecture-based formats also seem appreciated.

Awareness and Discovery Practices

Overall, awareness of existing digital infrastructures and resources via Huminfra and DARIAH appears to be very limited among respondents. Only 22% reported being familiar with the tools available via Huminfra, and just 11% with those provided through DARIAH. Similar patterns were observed for digital resources, with 21% aware of resources available via Huminfra and 10% via DARIAH. Awareness of training resources was even lower: only 15% were familiar with those offered through Huminfra and 6% with those from DARIAH. Furthermore, only 34 of the 209 respondents reported following Huminfra's or DARIAH's newsletters.

When asked how they typically learn about new digital tools and resources, the most common sources were colleagues (176), conferences (82), websites (81), and social media (60). Less common were email lists (47), newsletters (42), and online forums (31). These results indicate researchers primarily rely on personal networks and professional events rather than official communication channels to discover new tools, resources, and training opportunities.

Conclusion

The preliminary results of the survey reveal a community of researchers and cultural heritage professionals that is increasingly active in the use of digital tools and resources, yet still marked by uneven levels of competence, limited access to training, and low awareness of existing infrastructures.

The findings may be organized into five key areas:

Widespread use of digital tools and resources: Digital methods and resources are now integral to research and cultural heritage work in Sweden. This widespread use reflects the growing normalization of digital research practices; it highlights that digital engagement is no longer a niche activity but a shared element of scholarly and curatorial work.

A digital divide in proficiency: Despite this broad engagement, the survey reveals significant variation in digital competence. Roughly 40% of respondents rated their proficiency with digital tools relevant to the fields as modest or insufficient, indicating a digital divide between researchers with advanced technical skills and those with more limited experience.

Strong need and demand for training: A key finding is the high demand for structured training. Interest in training was strongest in areas such as data analysis, visualization, programming, and digitization, with a clear preference for hands-on, workshop-based formats.

Self-learning and informal discovery channels: The Swedish survey also shows that researchers primarily learn about new tools and methods through colleagues and conferences. This reliance on personal networks indicates that much digital competence is built through informal exchange rather than systematic training or infrastructure-mediated dissemination.

Low awareness of infrastructures: Finally, the survey highlights very low awareness of the tools, resources, and training available through digital infrastructures such as Huminfra and DARIAH-SE. This suggests a significant visibility gap and indicates that the infrastructures' potential is not yet fully realized within the Swedish research community.

References

Dallas, Costis, & Nephelie Chatzidiakou (eds.). European Survey on Scholarly Practices and Digital Needs in the Human Sciences. Athens: DARIAH-EU Digital Methods and Practices Observatory Working Group (DiMPO), 2022. https://doi.org/10.5281/zenodo.6583037

Session 2B — 10:30–12:10

10:30–10:50 *SHORT PAPER* **[36]**

Rule-based recognition of repetition

Ingerid Løyning Dale[1], Ranveig Kvinnsland[2]

[1] *National Library of Norway, Norway*
[2] *University of Oslo, Norway*

Keywords: *poetry analysis, computational linguistics, python*

This paper introduces poetry_analysis, a simple rule-based Python library to parse and annotate patterns of repetition in poetry and lyric texts, initially developed for Norwegian. Lyric features involving repetition present themselves in many forms, and poetry_analysis offers rule-based algorithms to extract patterns such as alliteration, anaphora, and end rhymes from texts where the newline is a meaningful separator.

These visual and positionally fixed patterns are often easy to see for the human eye but require strict definitions to operationalize so they can be detected and extracted algorithmically. The implementation within the poetry_analysis library defines the end rhyme as the overlap between the rhymes in the last stressed syllables on two verse lines within the same stanza, or a partial overlap between the last tokens on two verse lines. Alliteration is implemented as the repetition of a word-initial consonant in a word sequence where only function words are allowed to intervene between the alliterating words. Both end rhyme and alliteration are most often repetitions of sounds, and therefore the function for detecting these patterns can be used on phonemic transcriptions, if this is available for the user. Anaphora is a repetition of phrases and verse lines and is defined and recognised in two ways: 1) The repetition of a line-initial word or phrase across successive lines within the same stanza, or 2) the repetition of a phrase within the same verse line. This computational approach calls for considerably stricter definitions of these concepts than traditional literary analysis. However, this library is based on structures inherent to the lyric genre and can be used to explore repetition as a prominent lyric feature.

The library was first developed alongside, and to be used on, a corpus of Norwegian poetry from the 1890s (Kvinnsland et al. 2024). However, the step-by-step recognition of each pattern can be modified and parametrized by the user to fit specific needs and research requirements. The interface enables scaling annotations of lyric features in single lines, stanzas or poems, up to larger corpora of thousands of poems.

The initial lack of poetry corpora annotated with lyrical features in Norwegian meant that we had to rely on a more traditional rule based and test-driven code development of this annotation tool. A key limitation we have identified so far is the tools' inability to annotate examples of anaphora and alliteration where the repeating letters, words or phrases are more spaced out, but still repeated enough times to be "relevant" or "rhetorically effective". The tool may be used to create initial training data for machine learning models, which could then be used to annotate these "sparser" repetition patterns with less strict definitions and criteria in an attempt to get a higher recall of true anaphora and alliteration patterns. Until then, the unit tests in the tool's current version ensures that the tool annotates these patterns with a high precision.

Our preliminary evaluation on a test set of 100 individual poems, with 3159 verse lines, shows that the tool annotates end rhymes on orthographic text with an accuracy of 78.38%. For phonemic transcriptions the accuracy drops by ten percentage points to 68.38%, which is likely due to a mismatch between the writing standard in our test set and the writing standard that the transcription model was trained on (Røsok and Dale 2024).

The source code, documentation, and tutorials are open source and publicly available on Github, and the python package is easily installable from PyPi.

References

Kvinnsland, Ranveig, Ingerid Løyning Dale, and Lars Magne Tungland. 2024. 'Rediscovering the 1890s: A Norwegian Poetry Corpus'. In Proceedings of the Computational Humanities Research Conference 2024, edited by Wouter Haverals, Marijn Koolen, and Laure Thompson, vol. 3834. CEUR Workshop Proceedings. CEUR. https://ceur-ws.org/Vol-3834/#paper24.

Røsok, Marie Iversdatter, and Ingerid Løyning Dale. 2024. 'NB Uttale: A Norwegian Pronunciation Lexicon with Dialect Variation'. In Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), edited by Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue. ELRA and ICCL. https://aclanthology.org/2024.lrec-main.1056.

10:50–11:10 *SHORT PAPER* **[37]**

Lost in the Titles: Text-mining Metadata in the Digital Edition of Grundtvig's Works

Kirsten Vad, Katrine Laigaard Baunvig

Aarhus University, Denmark

Keywords: *Text-mining, metadata analysis, paratextual studies, N.F.S. Grundtvig, genre classification*

Grundtvig wrote thousands of texts – but what did he call them?

Across six decades, the influential Danish pastor, poet, and politician N.F.S. Grundtvig (1783–1872) produced sermons, hymns, poems, and a vast body of prose writings now being gathered in the digital scholarly edition *Grundtvig's Works* (scheduled for completion in 2030). While his writings have been extensively studied for their theological, linguistic, and cultural significance, the paratextual layer of his works – its titles – has received little critical attention. Yet titles serve as crucial thresholds to the texts: they mediate between author and reader, signal genre and interpretive frame, and reflect evolving publication practices (Genette 1997; Miller 1984). This paper explores what digital metadata can reveal about Grundtvig's authorship through a large-scale analysis of approximately 700 titles spanning 1804–1871.

While Franco Moretti's *"*Style, Inc.: Reflections on 7,000 Titles (British Novels, 1740–1850)*"* (2009) demonstrated how large-scale title analysis can illuminate transformations in literary form and book-market culture, our study extends this line of inquiry to a non-fictional, theological-literary corpus. By

examining Grundtvig's titles across sermons, hymns, and prose writings, we test the interpretive potential of title analysis beyond the novel tradition. Using text-mining methods including length analysis, word-frequency distributions, and part-of-speech tagging, we investigate how title structure and vocabulary vary across genres and over time. The analysis shows that certain genres demand more descriptive titles while others permit greater ambiguity. In treating titles as a distinct corpus of linguistic data, we adopt a distant-reading approach that reveals patterns invisible to traditional close-reading methods (Bode 2017). By analyzing titles and metadata across *Grundtvig's Works*, we uncover how titling strategies differ between lyric, prose, and composite genres. The dataset comprises 697 titles published between 1804 and 1871. On average, lyric titles are the briefest (3.6 words, 24 characters), often elliptical or dedicatory in tone; prose titles are the longest (5.1 words, 35 characters), reflecting the explanatory and argumentative mode of essays, sermons, and theological treatises; while titles in the 'other' category – dramas, periodicals, and collected works – occupy a middle ground (3.8 words, 26 characters) and tends to use titles that emphasize occasion, audience, or performance context, marking their circulation in public and institutional settings. These quantitative contrasts reveal genre-specific strategies of framing and communication that remain hidden within the infrastructural metadata of the digital edition (Bowman 2023; Liddle 2012). Treating titles as paratextual traces allows for a distant reading of Grundtvig's authorship that situates his writing practices within broader transformations of nineteenth-century publication culture (Baunvig 2021; Baunvig 2023).

Methodologically we situate the paper within current debates in digital hermeneutics and paratextual studies, demonstrating how structured editorial metadata, often treated as mere infrastructure, can become a productive site of interpretive inquiry (Bode 2017). We combine transparent, Python-based text-mining with exploratory analyses assisted by a large language model (GPT-5) (Papa 2025). The Python workflow provides reproducible measures of length, frequency, and distribution, while the LLM supports semantic clustering and interpretive description of title types (e.g. dedicatory, descriptive, theological). This hybrid approach addresses both the orthographic challenges of nineteenth-century Danish and the interpretive affordances of computational methods, demonstrating how computational precision and machine-assisted reading can complement each other. This reproducible workflow can be extended as the edition approaches completion in 2030, exemplifying how digital-humanities research can generate cumulative knowledge alongside evolving editorial projects.

While title analysis cannot capture authorial intention or reception history, it offers a reproducible, data-driven method for tracing formal and rhetorical shifts across a large-scale corpus. Our findings point to genre as a key variable in Grundtvig's self-presentation and to the potential of metadata as a bridge between editorial practice and literary interpretation. By re-examining a well-documented corpus through its infrastructural layer, we show how meaning can be recovered from what usually remains unseen – the metadata that structure access to canonical materials. In doing so, we demonstrate how digital methods can defamiliarize even thoroughly studied authors, opening new questions about the material and rhetorical dimensions of literary production.

References

Baunvig, Katrine Frøkjær (2021) "Fictional Realities of Modernity: The Fantastic Life of Demi-Goddess Dana in the Emerging Nation State of Denmark", in Mythology and Nation Building: N.F.S. Grundtvig and His European Contemporaries, edited by L. K. Martinsen, S. Bønding, and P.-B. Stahl, 97–134. Aarhus: Aarhus Universitetsforlag.

Baunvig, Katrine Frøkjær (2023) "'Each of Our Springs Has Lost Its Miraculous Power': The Range of a Religious Hotspot – A Distant Reading of Lourdes Representations in Denmark, 1858–1914", in Numen 70, no. 1: 43–69. https://doi.org/10.1163/15685276-12341675.

Bode, Katherine (2017) "The Equivalence of 'Close' and 'Distant' Reading; or, Toward a New Object for Data-Rich Literary History" in Modern Language Quarterly 78, no. 1: 77–106. https://doi.org/10.1215/00267929-3699787.

Bowman, Mike (2023) "Text-Mining Metadata: What Can Titles Tell Us of the History of Modern and Contemporary Art?", in Journal of Cultural Analytics 8, no. 1. https://doi.org/10.22148/001c.74602.

Genette, Gérard (1997) Paratexts: Thresholds of Interpretation. Translated by Jane E. Lewin. Cambridge: Cambridge University Press.

Liddle, Dallas (2012) "Reflections on 20,000 Victorian Newspapers: 'Distant Reading' The Times Using The Times Digital Archive", in Journal of Victorian Culture 17, no. 2: 230–37. https://doi.org/10.1080/13555502.2012.683151.

Miller, Carolyn R. (1984) "Genre as Social Action", in Quarterly Journal of Speech 70: 151-167.

Moretti, Franco (2009) "Style, Inc.: Reflections on 7,000 Titles (British Novels, 1740–1850)", in Critical Inquiry 36, no. 1: 134–58.

Papa, Erinda (2025) “(Re)Thinking Literary Interpretation in the Digital Age: AI, Virtual Reality, and Immersive Reading” in Open Journal of Social Sciences 13: 46–67. https://doi.org/10.4236/jss.2025.134003.

11:10–11:40 *LONG PAPER* **[38]**

Leveraging Large Language Models for Lemmatization and Translation of Finnic Runosongs

Lidia Pivovarova[1], Kati Kallio[2], Antti Kanner[3], Jakob Lindström[1], Eetu Mäkelä[1], Liina Saarlo[4], Kaarel Veskis[4], Mari Väina[4]

[1] *University of Helsinki, Finland*
[2] *Finnish Literature Society*
[3] *University of Turku*
[4] *Estonian Literary Museum*

Keywords: *Finnic runosongs, low-resource languages, lemmatization, large language models*

Finnic runosongs constitute a large multilingual oral poetic corpus collected between the sixteenth and twentieth centuries across Finnish, Karelian, Ingrian, Votic, and Estonian regions. The combined corpus, comprising about 250,000 texts, presents both an unprecedented resource and a methodological challenge for computational analysis. The material exhibits multiple dialects and orthographies, archaic morphology, and a high degree of poetic parallelism and repetition (Janicki et al. 2024). These characteristics make the corpus a representative example of non-canonical and low-resource heritage material: abundant but internally diverse and lacking uniform linguistic tools (Harvilahti 1992; Ross 2015; Saarinen 2018; Wiechetek et al. 2024; Paul et al. 2024).

The study examines to what extent large language models (LLMs) can assist in token-level linguistic analysis of runosongs. We focus on lemmatization—both in the original dialect and in a modern standard—and literal English translation. To provide a basis for evaluation, we created a manually annotated dataset of 206 songs (3,665 verses; 12,651 running words), sampled to ensure coverage across dialects, collection periods, and orthographic traditions. The corpus was manually annotated by folklorists specialized in Finnic runosongs. For each token, annotators supplied the normalized form, the lemma in the original dialect, the lemma in a modern standard (Finnish or Estonian) language, an etymological root corresponding to dictionary headwords, and a literal English translation. Annotators used reference grammars and dialect dictionaries but did not consult LLMs during the process.

The benchmarking combines manually curated data with a modular prompt architecture developed collectively by folklorists, linguists, and data scientists. Prompt design was an iterative process: domain specialists drafted linguistic and cultural context blocks describing dialectal features, poetic conventions, and common ambiguities, while technical partners added formatting constraints and retry logic to enforce stable outputs. Prompts were refined through small-scale qualitative tests before large-scale benchmarking. The final modular structure consists of three layers: a strict format layer enforcing complete tabular output, a linguistic context layer describing dialectal and poetic conventions, and a pipeline layer defining task sequence.

Six models were tested: five open (LLaMA 3.3–70B, Mixtral 8×22B, DBRX–Instruct, Llama-Poro 2–70B, and DeepSeek-R1) and one proprietary (Claude 3.7 Sonnet). Evaluation employed exact string match for categorical fields and cosine similarity for literal translations. The largest models performed best overall for lemma and root prediction, while smaller instruction-tuned models produced more consistent normalization. Claude reached around 75 percent accuracy for standard lemmas, DeepSeek about 65 percent, and the best open Finnish model (Poro) about 55 percent. This pattern suggests that over-generation may degrade orthographic accuracy even in high-capacity systems.

The comparison across prompt types yielded no consistent benefit from multi-stage or translation-first pipelines. Adding linguistic context improved results in some cases—most notably in Estonian etymological roots—but decreased accuracy elsewhere.

Qualitative analysis revealed typical model behaviors: confusion between normalization and standardization, substitution of rare lexical items by frequent synonyms, misinterpretation of homonymy and of archaic or poetic morphological endings, unstable definitions of the “root” field, and inconsistent treatment of refrains, onomatopoetic words, and culturally marked or obscene vocabulary. Such

phenomena illustrate both the potential and limits of applying LLMs to under-documented historical language material.

While the immediate aim is to benchmark current models, the broader goal is methodological. Runosongs exemplify the challenges of abundant but heterogeneous cultural data. Transparent, auditable LLM pipelines make it possible to process such corpora systematically while retaining linguistic structure. Future work will expand dialect coverage, integrate lexicon-based prompting using etymological and dialectal dictionaries, and test whether token-level annotations enable further analyses, such as motif clustering or regional variation mapping.

This study confirms that LLMs can meaningfully support linguistic expertise in the analysis of complex poetic heritage. Their output remains sensitive to prompt structure and domain context. The combination of human expertise, reproducible benchmarks, and explicit evaluation offers a path toward fruitful application of language models to non-standard and non-canonical materials. Even the best models make mistakes, and some parts of the material are simply hard to interpret, even for human specialists. Our approach is to recognize these limits, quantify them, and use computational methods to support rather than replace close linguistic and cultural interpretation.

References

Harvilahti, Lauri. 1992. Kertovan runon keinot. Inkeriläisen runoepiikan tuottamisesta. Helsinki: SKS.

Janicki, Maciej, Eetu Mäkelä, Mari Väina, and Kati Kallio. 2024. “Developing a Digital Research Environment for Finnic Oral Poetry.” Baltic Journal of Modern Computing 12 (4): 535–547. https://doi.org/10.22364/bjmc.2024.12.4.15.

Paul, Ronny et al. 2024. “Towards a More Inclusive AI: Progress and Perspectives in Large Language Model Training for the Sámi Language.” arXiv preprint 2405.05777. https://arxiv.org/pdf/2405.05777.

Ross, Kristiina. 2015. “Regivärsist kirikulauluni: Kuidas ja milleks kõrvutada vanu allkeeli [From runo verse to hymns: How and why compare old sublanguages].” Keel ja Kirjandus 7: 457–470. https://doi.org/10.54013/kk692a1.

Saarinen, Jukka. 2018. Runolaulun poetiikka: Säe, syntaksi ja parallelismi Arhippa Perttusen runoissa. Helsinki: Helsingin yliopisto. http://urn.fi/URN:ISBN:978-951-51-3919-1.

Wiechetek, Linda, Flammie A. Pirinen, Børre Gaup, Trond Trosterud, Maja Lisa Kappfjell, and Sjur Moshagen. 2024. “The Ethical Question – Use of Indigenous Corpora for Large Language Models.” In Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), 15922–15931. Torino: ELRA / ICCL. https://aclanthology.org/2024.lrec-main.1383/.

11:40–12:10 *LONG PAPER* **[39]**

Computationally Identifying Recurrent Units in Finnic Oral Poetry

Eetu Mäkelä[1], Kati Kallio[2], Mari Väina[3], Liina Saarlo[3], Jakob Lindström[1], Venla Sykäri[2], Antti Kanner[4], Maciej Janicki[5], Lidia Pivovarova[1], Kaarel Veskis[2]

[1] *University of Helsinki, Finland*
[2] *Finnish Literature Society*
[3] *Estonian Literary Museum*
[4] *University of Turku*
[5] *CSC – IT Center for Science*

Keywords: *oral poetry, recurrent units, variation, complex data, pattern discovery*

Creativity in traditional oral poetry works differently from, for example, literary tradition or contemporary popular music. Variation is an inherent feature of folklore, meaning that in each performance the text may be reorganised, adapted to the circumstances and the audience. Some elements are replaced in a new performance, while others are more stable. This variation is versatile, flexible and multilevel, and it typically takes different forms, levels and degrees in different languages, local cultures, singers, performances, genres and song types.

In Finnic oral poetry (runosongs, Kalevalaic poetry), many levels and types of recurrent structures can be identified. These range from bare recurring filler words, words commonly appearing in similar metrical positions, short multi-word formulae (poetically and metrically motivated, recurring collocations) and verse types to various ways in which verses or groups of verses can recur together, such as parallel

sections, poetic chains, motifs, chains of motifs (plots or established sequences), poem types or combinations of these. Some elements tend to appear in one poetic context only, while others can be multi-use, multi-purpose ones. (See e.g. Harend 2024; Harvilahti 1992; Saarinen 2018, Saarlo 2005).

The analysis of all these levels of recurrence is made complicated by the fact that mostly none of the recurrences are rigid, but each permits an amount of variation. For example, in different occurrences of stereotypical verses, individual words can vary, verse pairs or groups can often be selected from a larger set of options or can be extended by more random verses. For the larger recurring units, there is even more possibility for variation, with, for example, unrelated verses and sections sometimes appearing in the middle of a more-or-less fixed and established plot. Motifs again are often not expressed through stable verbal forms but are recognised semantically through recurring meanings and narrative functions. Their identification depends on subtle textual cues rather than fixed formal features, which makes them resistant to strict definition. Further, with all of these levels of variation occurring simultaneously, it is immensely difficult to extract patterns from a particular level from all the confounding noise caused by variations on the other levels.

All of the above make the computational identification and analysis of recurrent patterns in oral poetry extremely difficult. In this presentation, we will discuss the multiple ways we've approached the problem within the long-running FILTER collaboration (see Janicki 2022; 2023; Janicki et al. 2024a; 2024b; Kallio et al. 2024; Sarv & Järv 2023; Sarv et al. 2021; Sarv et al. 2024, Väina (forthcoming)). Key highlights include how we've used information from particular levels to reach others. For example, we've combined information gained from computational (syntactic or semantic) verse clustering with manually curated poem type information to extract archetypal verses signifying the presence of a particular poem type, as well as stereotypical generic verses that easily cross types.

We also contrast our explorations into bottom-up and top-down approaches. In the bottom-up approach, we start with regularly co-occurring verses to build more complex recurring patterns, while in the top-down approach, we start with complete recorded poems exhibiting a particular poem type, and align, cluster or otherwise divide them to identify the major variants. In our presentation, we will analyse how these approaches provide complementary avenues to get at the complex intermediate-sized recurring structures in between these two endpoints. Finally, we will discuss our experiences in how far one is able to get with computational approaches, and where manual interventions are still required.

References

Foley, John Miles 1988: The theory of oral composition: History and methodology. Bloomington: Indiana University Press.

Helina Harend 2024. Ema-, isa-, õe- ja vennanimetused eesti regilauludes. Master's thesis. University of Tartu, Faculty of Arts and Humanities, Institute of Estonian and General Linguistics. https://hdl.handle.net/10062/102022

Janicki, Maciej, Eetu Mäkelä, Mari Väina ja Kati Kallio. 2024a. Developing a Digital Research Environment for Finnic Oral Poetry. Baltic Journal of Modern Computing 12(4), 535–547. https://doi.org/10.22364/bjmc.2024.12.4.15.

Janicki, Maciej & Kati Kallio & Mari Sarv & Eetu Mäkelä 2024b: Distributional criteria for identifying formulas in Finnic oral poetry. In Formulaic Language in Historical Research and Data Extraction. International Institute for Social History, Amsterdam, 7.-9.02.2024. Ed. Marijn Koolen. Amsterdam: Huygens Institute for History and Culture of the Netherlands, Royal Netherlands Academy of Arts and Sciences. Zenodo, 1−17. https://zenodo.org/doi/10.5281/zenodo.10478324.

Janicki, Maciej 2023. Large-scale weighted sequence alignment for the study of intertextuality in Finnic oral folk poetry. In: Journal of Data Mining and Digital Humanities, NLP4DH. https://doi.org/10.46298/jdmdh.11390

Janicki, Maciej 2022: Optimizing the weighted sequence alignment algorithm for large-scale text similarity computation. In M. Hämäläinen, K. Alnajjar, N. Partanen, & J. Rueter (Eds.), Proceedings of the 2nd International Workshop on Natural Language Processing for Digital Humanities, pp. 96–100. https://aclanthology.org/2022.nlp4dh-1.13/

Kallio, Kati, Mari Väina, Maciej Janicki ja Eetu Mäkelä. 2024. "Bridging Northern and Southern Traditions in the Finnic Corpus of Oral Poetry." Folklore: Electronic Journal of Folklore 94: 191–232. https://www.folklore.ee/folklore/vol94/finnic_corpus.pdf https://doi.org/10.7592/FEJF2024.94.finnic_corpus

Lord, Albert B. 1960. The Singer of Tales. Cambridge: Harvard University Press. [Repr. 2000.]

Saarinen, Jukka 2018: Runolaulun poetiikka: Säe, syntaksi ja parallelismi Arhippa Perttusen runoissa. Helsinki: Helsingin yliopisto. http://urn.fi/URN:ISBN:ISBN 978-951-51-3919-1 (PDF).

Harvilahti, Lauri. 1992. Kertovan runon keinot. Inkeriläisen runoepiikan tuottamisesta. Helsinki: SKS.

Sarv, Mari & Kati Kallio & Maciej M. Janicki 2024: Arvutuslikke vaateid läänemeresoome regilaulude varieeruvusele: “Harja otsimine” ja “Mõõk merest”. Keel ja Kirjandus, 67(3), 238–259. https://doi.org/10.54013/kk795a2

Saarlo, Liina. 2005. Eesti regilaulude stereotüüpiast. Tartu: Tartu Ülikooli kirjastus. http://hdl.handle.net/10062/838

Sarv, Mari; Järv, Risto 2023. Layers of Folkloric Variation: Computational Explorations of Poetic and Narrative text Corpora. Folklore: Electronic Journal of Folklore, 90, 233−266. DOI: 10.7592/FEJF2023.90.s, https://www.folklore.ee/folklore/vol90/sarv_jarv.pdf

Sarv, Mari; Kallio, Kati; Janicki, Maciej; Mäkelä, Eetu (2021). Metric variation in the Finnic runosong tradition: A Rough Computational Analysis of the Multilingual Corpus. In: Petr Plecháč, Robert Kolár, Anne-Sophie Bories, Jakub Říha (Ed.). Tackling the Toolkit. Plotting Poetry through Computational Literary Studies (131−150). Prague: Institute of Czech Literature CAS. DOI: 10.51305/ICL.CZ.9788076580336.09.

Sarv, Mari; Kallio, Kati; Janicki, Maciej (2024). Arvutuslikke vaateid läänemeresoome regilaulude varieeruvusele: "Harja otsimine" ja "Mõõk merest". Keel Ja Kirjandus, LXVII (3), 238−259. DOI: 10.54013/kk795a2.

Väina, Mari (forthcoming). Regional variation of Finnic runosongs on the basis of word-frequencies. In: M. Väina, T. Särg (Eds.). Expressions and Impressions: Personal and Communal Aspects of Traditional Singing. (139−156). Helsinki: Finnish Literary Society. (Studia Fennica Folkloristica).

Session 2C — 10:30–12:30

10:30–10:50 **Consistency checking in a cloud of interlinked Cultural Heritage knowledge graphs – first results of using the SampoSampo data service and portal**
Petri Leskinen, Annastiina Ahola, Heikki Rantala, Jouni Tuominen, Eero Hyvönen

10:50–11:10 **Between Print and Digital: Data Sharing and Community Formation in Digital Humanities Pedagogy before the Web**
Sofia Papastamkou

11:10–11:30 **ARCH-ON: A new ontological framework to describe archaeological objects for Digital Humanities research**
Pieterjan Deckers, Eero Hyvönen, Michael Lewis, Eljas Oksanen, Heikki Rantala, Jouni Tuominen

11:30–12:00 **A Goldmine Unknown: Mapping and Visualizing the Saga Archive though Digital Methods**
Åsa Warnqvist, Julia Beck

12:00–12:30 **Hidden Catastrophes: Mapping Disaster Narratives in Non-Canonical Nordic and Baltic Children’s Literature and Fairy Tales Through Digital Archives**
Olena Koliasa

10:30–10:50 *SHORT PAPER* **[40]**

Consistency checking in a cloud of interlinked Cultural Heritage knowledge graphs – first results of using the SampoSampo data service and portal

Petri Leskinen[1,2], Annastiina Ahola[1], Heikki Rantala[1], Jouni Tuominen[2,1], Eero Hyvönen[1]
[1] *Aalto University, Finland*
[2] *University of Helsinki, Finland*

Keywords: *linked data, digital humanities, entity alignment, data validation, semantic portal, data analysis, knowledge discovery*

This paper presents a novel use case and functionalities of the new data linking service and portal "SampoSampo — Connecting Everything to Everything Else" on top of it, based on a Linked Open Data (LOD) cloud of related Cultural Heritage (CH) knowledge graphs (KG) of different application domains. This new Sampo system interlinks data about historical Finnish people, organizations, places, and events from 11 earlier Sampo systems in use on the Web as well as 8 other Finnish and international sources of data on the Web. This paper focuses on one particular use case of SampoSampo: how to detect automatically conflicting and complementary data about entities in an interlinked cloud of CH KGs. With provenience information attached, this is useful 1) for data publishers for checking the quality and possible errors in their data by comparing it with datasets provided by the other publishers, and 2)

for data consumers for verifying the quality of data based on multiple primary sources. Our first results show some striking conflicts and disagreements between the primary data sources interlinked in SampoSampo.

10:50–11:10 *SHORT PAPER* **[41]**

Between Print and Digital: Data Sharing and Community Formation in Digital Humanities Pedagogy before the Web

Sofia Papastamkou

C2DH, University of Luxembourg, Luxembourg

Keywords: *data sharing, digital humanities, digital history*

Focusing on the *Computers and the Humanities* surveys (1971–1987), this paper explores how early humanities computing pedagogy negotiated the scarcity of digital infrastructures and anticipated today's culture of abundance in data sharing and open pedagogy.

Between 1971 and 1987, the journal *Computers and the Humanities* (CHum) organized five surveys on teaching computers in the humanities (Bowles 1971, Campo 1972, Allen 1974, Rudman 1978, Rudman 1987). The journal itself—founded in 1966 by Joseph Raben, professor of English at Queens College (CUNY), with support from an IBM grant—was one of the signs that a community around humanities computing was beginning to take shape. The surveys it organized served as a means of collecting and disseminating information on teaching with computers in the humanities, with the aim of further inspiring such courses. Contributing to the institutionalization of these courses and, ultimately, strengthening a community of practice in humanities computing was among the programmatic aims of the journal (Raben 1966).

The five surveys constitute an important source for exploring the place of pedagogy in the history of digital history and the digital humanities (Papastamkou 2024)—an area that remains relatively understudied in the field's literature (Georgopoulou *et al*. 2025, Hirsch 2012). At the same time, the surveys serve as a case study through which one can observe the evolution of data-sharing needs and practices in the humanistic scholarly community, before the digital data deluge of the past two decades and the subsequent development of related infrastructures and institutional policies.

The surveys took place in the following years: 1971, 1972, 1974, 1978, and 1987. In each case, the results were analyzed and presented in a dedicated article published in the journal. The data were also included in the corresponding article as appendices, which contained both primary (raw) and secondary (processed, calculated, or synthesized) data. Thus, the editorial presentation of the survey data and their scientific interpretation appeared unified within a single publication format—the printed scholarly article. This was neither unique nor new (Plutniak 2023). However, from the first to the last survey, a process of *autonomization* of the editorial objects (data, article) seems to have been underway. Beginning with the 1978 survey, the volume of data increased significantly. This was also the last survey to publish primary data within the journal article. The 1987 publication contained only secondary data in its appendices because the volume of the survey data now made it impossible to include them in the print article.

At the same time, both the need and the desire to share data independently of the article became increasingly evident. This led to the proposal of creating a data clearinghouse to distribute the datasets via floppy disk (Rudman 1987). Although this clearinghouse was never realized, the data were informally shared with anyone who requested them—until their eventual loss in a fire (Rudman 2025). It has since been possible to partly reconstruct the survey dataset by extracting the data from the printed appendices of the first four surveys. Nonetheless, the source remains fragmentary: except for one appendix, the 1987 survey is accessible only through secondary data and the elements discussed in the accompanying analysis (Papastamkou 2024 and 2025).

Through this specific case, the late 1980s appear as a moment when several constitutive elements of sharing—today considered a defining feature of the digital humanities—were already in place: a community, shared domains of interest (in this case, humanities computing pedagogy), and available information in the form of data. What was still missing were the distributed digital infrastructures needed to sustain these practices. This paper focuses on that print/digital *in-between* moment, which preceded

the community-led and open pedagogical practices of the 2000s onward (Kirschenbaum 2010; Varin 2013). These later developments fully supported pedagogical outputs that were electronically published and openly distributed through web technologies—what has been described as the “invisible college of digital history (Crymble 2019).

References

Bowles, Edmund A. “Towards a Computer Curriculum for the Humanities.” Computers and the Humanities 6, no. 1 (1971): 35–38. https://doi.org/10.1007/BF02402323.

Campo, Leila de. “Computer Courses for the Humanist: A Survey.” Computers and the Humanities 7, no. 1 (1972): 57–62. https://doi.org/10.1007/BF02403762.

Crymble, Adam. Technology and the Historian: Transformations in the Digital Age. Topics in the Digital Humanities. University of Illinois Press, 2021.

Hirsch, Brett D., ed. Digital Humanities Pedagogy : Practices, Principles and Politics. In Digital Humanities Pedagogy : Practices, Principles and Politics. Digital Humanities Series. Open Book Publishers, 2012. https://books.openedition.org/obp/1605.

Kirschenbaum, Matthew. “What Is Digital Humanities and What’s It Doing in English Departments?”. In Debates in the Digital Humanities. University of Minnesota Press, 2018.

Papastamkou, Sofia. “Teaching Historians “the ways of the machine”: Proto-debates, Actors, and Practices on Code Literacy in the Humanities, 1966 -1987.” Paper presented at Revolutionary, Disruptive, or Just Repeating Itself? Tracing the History of Digital History” #dhiha9, Paris, France, 24 October 2024 (to be published in 2026)

Papastamkou, Sofia. “A Digital Literacy in the Making. Teaching Computers to Historians/Humanists, 1960s-1980s.” Paper presented at History of Knowledhe Conference, 8-10 October 2025, LUCK Lund Centre for the History of Knowledge

Plutniak, Sébastien. “L’autonomisation éditoriale de la publication des données. Des tirés-à-part aux data journals en archéologie (1950-2000).” In Norbert Verdier; Hélène Védrine; Alexia Kalantzis. Les

périodiques comme médiateurs culturels. Autour de la diffusion des savoirs. Séminaire PéLiAS

(Périodiques, Littérature, Arts, Sciences) 2019-2022, MSH Paris-Saclay Éditions, pp.155-173, 2023

Raben, Joseph. “Prospect.” Computers and the Humanities 1, no. 1 (1966): 1–2.

Rudman, Joseph. “Computer Courses for Humanists: A Survey.” Computers and the Humanities 12, no. 3 (1978): 253–79. https://doi.org/10.1007/BF02400087.

Rudman, Joseph. “Teaching Computers and the Humanities Courses: A Survey.” Computers and the Humanities 21, no. 4 (1987): 235–43. https://doi.org/10.1007/BF00517812

Varin, Vanessa, “A Thoughtful Retrospective of One Historian’s Experience at THATCamp”, AHA Today January 11, 2013 available at Perspectives at History https://www.historians.org/, https://www.historians.org/research-and-publications/perspectives-on-history/january-2013/a-thoughtful-retrospective-of-one-historians-experience-at-thatcamp

11:10–11:30 *SHORT PAPER* **[42]**

ARCH-ON: A new ontological framework to describe archaeological objects for Digital Humanities research

Pieterjan Deckers[1], Eero Hyvönen[2,3], Michael Lewis[4,3,5], Eljas Oksanen[5,3], Heikki Rantala[2], Jouni Tuominen[3,2]

[1] *KU Leuven, Belgium*
[2] *Aalto University, Finland*
[3] *University of Helsinki, Finland*
[4] *British Museum, UK*
[5] *University of Reading, UK*

Keywords: *archaeology and history of religion, data harmonization, Europe, Linked Open Data, ontologies*

This paper presents the idea and preliminary results of the ARCH-ON (Archaeological Ontologies) pilot project[1]. This project aims to lay the groundwork for the semantic description and classification of archaeological artefacts, enabling new approaches to the recording, study, and public accessibility of information regarding material culture from the past.

[1] ARCH-ON project homepage: https://seco.cs.aalto.fi/projects/arch-on/

11:30–12:00 *LONG PAPER* **[43]**

A Goldmine Unknown: Mapping and Visualizing the Saga Archive though Digital Methods

Åsa Warnqvist[1], Julia Beck[2]

[1] *The Swedish Institute for Children's Books, Sweden*

[2] *Gothenburg University, Sweden*

Keywords: *children's literature, publishing industry, archives, digital archiving, data modeling*

Svensk läraretidnings förlag, the Swedish Teachers' Magazine's Publishing House, was one of the first Swedish publishers to focus primarily on publications aimed at children and youth. Publishing books, magazines, and other items aimed at children from 1896 to the mid-twentieth century, the publishing house performed one of the most ambitious and long-running reading-promotion projects ever undertaken in Sweden. It had a huge impact on the development of children's literature in Sweden in the beginning of the twentieth century (see, e.g., Klingberg). The initiative began as a collaboration between teachers, and the aim was to promote reading at school and at home by providing cheap, quality literature for all children, including those with limited means. It was hugely successful.

The company records from the publisher are kept in the Saga Archive housed at the Swedish Institute for Children's Books in Stockholm (SBI). The archive contains more than twenty shelf meters of archival material, plus almost all the published books, journals and other publications. The archive is extensive but has received little scholarly attention.

In 2022, SBI received funding from Riksbankens Jubileumsfond, the National Bank's Jubilee Fund, to embark on a project to make the Saga Archive more available. The main purpose of the project is to explore, map, and digitize material from the archive. This aim is realized through creating a database including digital bibliographies of the publishing house's publications and scans of a selection of the original documents, images, and publications, to make metadata searchable and material available online. The project is conducted in collaboration with the Gothenburg Research Infrastructure in Digital Humanities (GRIDH).

In this paper, we discuss the challenges that the project group of researchers, archivists, and research engineers have encountered when analyzing and mapping the archival material, and the choices we have made. The large number of books, magazines, and other items published by the Swedish Teachers' Magazine's Publishing House, and the fact that the editorial team as a rule revised and/or re-used material in different kinds of publications, create links between different documents and sections of the physical archive. Often these links are difficult to detect due to the vast number of items. The many links proved a challenge when defining the entities, attributes, structures, and hierarchies of the material, and the relationships between different items or groups of items, but at the same time it provided a possibility to bring scattered material together through digital methods. We have used Omeka-S as the web-publishing platform for the organization and display of the material, and the nature of the material has required a number of adaptations and adjustments to the application. How can a database be tailored to represent a publishing house of this kind and its rich and dispersed archive? How can metadata be modeled to ensure logical and consistent organization over time? These are the main questions guiding our presentation of the work we have conducted so far.

Digital methodologies promise to have a significant impact on multiple areas of children's literature, and the possibility of new methodological approaches is frequently stressed in the conceptualization of digital archives (see, amongst others, Dicinoski; Hui; Morgan et al). A more troubling aspect, however, is the threat of the disappearance of context. Claire Brennan warns that the facility of searching online can strip sources of context and provenance. For example, as Shafquat Towheed notes, the specificity of reading is difficult to reconstruct from digital collections (see also Escobar Varela). Towheed formulates what we want to term the "digital availability paradox": "Paradoxically, while the digitization of nineteenth-century newspapers means that they have never been more widely available, the very experience of casually browsing through the pages of an essentially disposable publication has become increasingly remote and difficult to reconstruct" (142).

The aim of the digital Saga Archive is to create a platform that overcomes this digital availability paradox and provides researchers not only with digitized documents but also describes the relationships

between these documents and connections between actors, such as key persons, documents, publications, editing practices, manuscripts, letters, and other materials. The project group does not mainly view context as a preexisting or fixed historical, sociological, or pedagogical discourse surrounding the publishing house. Rather, context becomes visible in the systematic linking between different material within the digital archive, creating networks of relationships that can stimulate the researcher to a context-driven exploration of the documents. The scientific creation of a rich digital web of relationships between actors as such, can display in new ways the factual material preserved in the brick-and-mortar archive.

In addition, documents in the digital Saga Archive are supplemented with systematic and nuanced sets of metadata explaining the status of the documents. The original works are heavily adapted by the editors for the presumed audience, addressing aesthetic and pedagogical ideals that change over time. Many of the preserved manuscripts are filled with remarks, corrections, and suggestions on different levels. Making the annotations and marginalia in the documents digitally available, with metadata and networks of relationships, can contribute to a new and contextually enriched understanding of the production, publishing, and distribution of children's literature. Additionally, the connections can be contextualized by using reification (Dodds et al) in Omeka-S that allows annotating statements. Moreover, in order to create metadata that is interoperable with external data sources, the project group makes use of linked open data resources from *Wikidata*, *LIBRIS* (the Swedish union catalog), and VIAF for persons and organizations. We will also expose our database entities and assign permanent identifiers so that external users can link to, and retrieve data from, our database.

The vast amount and kinds of archival material constitute a valuable resource. The project has found ways of describing and digitizing the extensive archive material and made the artifacts previously hard to survey interact. We are building the database based on linked data principles (WWW Consortium) using structured data about the publishing of the publishing house and the structure and content of the archive. After the release of the online application in May 2026, there will be opportunities to explore and search both the complex publishing process (through entries like title, author, type of publication, and book series), the archive (through archive headings and content of archive volumes such as minutes, letters, and manuscripts), and the relation between the two—in other words, clarifying what archive material precedes a published book or magazine.

The comprehensive view of this project includes making the processes and mechanisms behind the production and distribution of the books visible beyond simply the finished product, which connects the project to the field of digital genetic criticism (Hay). However, the archive first and foremost makes it possible to study the publishing house's processes and rhetoric of editing (Widhe), rather than only the author's creative process. This means the publishing house's work can be studied in ways which have not been possible before. Thanks to the wide distribution and influence of the publishing house's publications, the archive sheds light on the development of Swedish children's literature in general during the period between the 1890s and the 1970s, as well as changing views on culture and cultural awareness. It also highlights the cultural radicalism of its time through the wish to make sure that all children, not only children in rich families, were given access to literature. The archival material illustrates how the editors, at different times, adapted the published text pedagogically, psychologically, and aesthetically to the prevailing view of children's needs and resources. The project enables resources that together describe intertextual relationships and literary processes, which facilitate network analyses of the people and actors involved in the publication of the works and their positions. This, in turn, enables an analysis of the publisher's position in the literary field at large and its importance to the development of the Swedish twentieth-century book market industry.

The digital Saga Archive not only offers an opportunity to study publication processes but also to further develop existing theories on the interplay between different value systems within literature, canon formation, mediation of literature, and reception theory. This new infrastructure will make it possible to study the role and changing conditions of children's literature, the dissemination of literature, and the reading of literature on both micro and macro levels in new ways. Through the digitization of the archive, research is facilitated on a period of Swedish history that is characterized by major societal changes and renegotiations on the roles of children and childhood. Moreover, the digital Saga Archive offers a range of research opportunities within a variety of fields – literature, art and book history, translation, education, gender studies, social history, and a range of other areas. Digitization and searchability in the tailored online resource will provide visible, accessible, and user-friendly material in a way that it has not been before.

References

Brennan, Claire. “Digital Humanities, Digital Methods, Digital History, and Digital Outputs: History Writing and the Digital Revolution.” History Compass, vol. 16, 2018, pp. 1–12.

Dicinoski, Michelle. “Digital Archives and Cultural Memory: Discovering Lost Histories in Digitised Australian Children’s Literature 1851–1945.” Papers: Explorations into Children’s Literature, vol. 22, no. 1, 2012, pp. 110–20. https://search.informit.org/doi/10.3316/informit.798785086607125.

Dodds, Leagh; Davis, Ian. “Linked Data Patterns: A Pattern Catalogue for Modelling, Publishing, and Consuming Linked Data”. 2022. https://patterns.dataincubator.org/

Escobar Varela, Miguel. “The Archive as Repertoire: Transience and Sustainability in Digital Archives.” DHQ: Digital Humanities Quarterly, vol. 10, no. 4, 2016, pp. 1–8.

Hay, Louis. “Genetic Criticism: Another Approach to Writing?” Research on Writing: Multiple Perspectives, edited by Sylvie Plane et al, WAC Clearinghouse, 2017, pp. 531–47. https://doi.org/10.37514/INT-B.2017.0919.

Hui, Haifeng. “What Can Digital Humanities Do for Literary Adaptation Studies: Distant Reading of Children’s Editions of Robinson Crusoe.” Digital Scholarship in the Humanities, vol. 38, no. 4, 2023, pp. 1564–76. https://doi.org/10.1093/llc/fqad059.

Klingberg, Göte. Sekelskiftets barnbokssyn och Barnbiblioteket Saga [Turn-of-the-century Views on Children’s Books and the Children’s Library Saga]. Svensk läraretidnings förlag, 1966.

Morgan, Marina, et al. “Digital Humanities and Metadata: Linking the Past to the Digital Future.” International Conference on Dublin Core and Metadata Applications, 2013, pp. 206–8. https://dcpapers.dublincore.org/pubs/article/view/3696.

Towheed, Shafquat. “Reading in the Digital Archive.” Journal of Victorian Culture, vol. 15, no. 1, 2010, pp. 139–43.

Widhe, Olle. “Barnbiblioteket Saga och den unga läsaren: Retoriska manuskriptpraktiker på en barnboksredaktion” [“The Children’s Library Saga and the Young Reader: The Rhetoric of Editing in Children’s Literature Publishing].” Barnboken: Journal of Children’s Literature Research, vol. 46, 2023, pp. 1–26. http://dx.doi.org/10.14811/clr.v46.819.

World Wide Web Consortium. “Best Practices for Publishing Linked Data ". 2014. https://www.w3.org/TR/ld-bp/

12:00–12:30 *LONG PAPER* **[44]**

Hidden Catastrophes: Mapping Disaster Narratives in Non-Canonical Nordic and Baltic Children’s Literature and Fairy Tales Through Digital Archives

Olena Koliasa[1,2]

[1] *Sydney University, Australia*

[2] *Lviv Ivan Franko National University*

Keywords: *disaster, narratives, catastrophe, children’s literature, fairy tales*

Introduction

This research examines how disaster narratives permeate children’s literature and fairy tales through a computational analysis of overlooked digital collections, revealing the cultural anxieties and coping mechanisms embedded within seemingly innocent texts. While canonical children’s literature scholarship has focused on well-known tales and major publishers, vast digitised archives contain thousands of marginalised children’s books, regional fairy tale variants, educational pamphlets, and ephemeral publications that encode complex disaster discourse within accessible narratives. The Nordic and Baltic regions offer a vibrant context for this investigation, given their shared experiences of historical catastrophes – from medieval plagues and famines to twentieth-century wars, occupations, and environmental disasters. These traumatic events have left indelible marks on cultural memory, often finding expression through children’s literature as societies attempt to transmit essential survival knowledge while protecting young minds from overwhelming realities. The abundance of digitised materials across Nordic and Baltic cultural institutions creates unprecedented opportunities to trace these patterns computationally across vast textual corpora.

Children’s literature serves as a unique cultural repository where societies process collective trauma, environmental threats, and social catastrophes through metaphor, allegory, and simplified narrative structures. Unlike adult literature, children’s texts must navigate the complex pedagogical challenge of preparing young audiences for potential disasters while maintaining psychological safety and hope. This dual function creates distinctive linguistic and narrative patterns that reveal how cultures conceptualise catastrophe, resilience, and recovery across generational boundaries. The concept of

DISASTER in this research extends beyond natural catastrophes to encompass cultural, linguistic, and social disruptions that reshape communities and identities. War, language suppression, religious persecution, and environmental degradation all constitute forms of catastrophe that Nordic and Baltic societies have processed through children's literature. These texts often preserve the cultural memory of traumatic events that official histories might minimise or suppress, making them invaluable sources for understanding how marginalised communities experienced and interpreted disasters.

Fairy tales, in particular, serve as repositories of ancient wisdom about disasters, encoding survival strategies and moral frameworks that have been developed over centuries of catastrophic experience. Regional variants of international tale types often reflect local environmental hazards, historical traumas, and cultural responses to crisis. The digitisation of previously scattered regional collections now enables large-scale comparative analysis of how disaster motifs evolve across cultural and linguistic boundaries within the Nordic-Baltic region.

Research Questions

This research investigates the following core question: How do disaster narratives in non-canonical Nordic-Baltic children's literature differ from those in canonical texts in terms of linguistic patterns, narrative structures, and thematic content? What distinctive linguistic and narrative patterns characterise disaster discourse across different Nordic-Baltic cultures, and how do these patterns reflect regional historical experiences and cultural values? How have disaster motifs in children's literature evolved from medieval periods through contemporary times, and what historical events correlate with shifts in disaster representation?

Materials

The primary digital corpora will include materials from the following institutional repositories: Royal Danish Library's Digital Collections (historical children's books, educational materials, and periodicals), National Library of Sweden's Digital Collections (regional publications, fairy tale collections, and wartime materials), National Library of Norway's Digital Archive (Children's literature spanning medieval to contemporary periods), National Library of Finland (Multilingual children's texts and educational pamphlets), Estonian Folklore Archives (Regional fairy tale variants and oral tradition transcriptions), National Libraries of Latvia and Lithuania (Post-occupation children's literature and cultural revival materials).

These institutional repositories contain thousands of digitised children's books, educational materials, and regional publications that remain largely unexplored by disaster studies scholars. Specific types of materials include marginalised children's books from regional publishers, fairy tale variants and folklore collections, educational pamphlets and textbooks, wartime children's publications, post-disaster educational materials, translated texts revealing cross-cultural transmission, and ephemeral publications (such as pamphlets, periodicals, and primers). The corpus will encompass materials in Danish, Swedish, Norwegian, Finnish, Estonian, Latvian, and Lithuanian, with selected texts in minority languages where available.

Methodology

This research employs topic modelling, sentiment analysis, and comparative textual analysis across Nordic and Baltic digitised children's collections to uncover patterns in how disaster – from natural catastrophes to war, disease, and social upheaval – is linguistically and culturally transmitted to young audiences through non-canonical sources. The methodology combines distant reading of extensive digital corpora with close analysis of specific textual variants to trace the evolution of disaster motifs in children's literature.

Multilingual Topic Modelling. To identify disaster-related themes across different languages and cultural contexts, utilising Latent Dirichlet Allocation (LDA) and other probabilistic models adapted for multilingual corpora. This will reveal thematic clusters related to specific disaster types (war, famine, plague, environmental catastrophe) and their regional variations.

Sentiment Analysis. To trace emotional registers associated with catastrophic narratives, examining how fear, hope, resilience, and despair are linguistically encoded in age-appropriate discourse. This analysis will reveal how different cultures strike a balance between psychological protection and disaster preparedness.

Network Analysis. To map relationships between characters, events, and moral frameworks in disaster-themed children's texts. This will illuminate narrative structures that connect protagonists, antagonists, helping figures, and catastrophic events within culturally specific moral universes.

Named Entity Recognition (NER). To identify references to specific historical disasters, geographical locations, and temporal markers that anchor fictional narratives to historical events. This will enable correlation analysis between documented disasters and literary responses.

Stylometric Analysis. To reveal how disaster discourse adapts linguistically for child audiences across different Nordic and Baltic cultures, examining vocabulary simplification, sentence structure, and rhetorical strategies specific to pedagogical contexts.

Analytical Framework

The research will employ both synchronic and diachronic approaches, analysing how disaster narratives function within specific historical moments while tracing their evolution across time periods. Particular attention will be paid to wartime children's publications, post-disaster educational materials, and translated texts that reveal cross-cultural transmission of disaster narratives within the region.

The analytical framework will examine disaster narratives at multiple levels: *lexical analysis* of disaster terminology and metaphor systems, *narrative structure analysis* of catastrophe-recovery patterns, *character analysis* of disaster protagonists and their coping strategies, and *cultural analysis* of moral frameworks surrounding disaster response. This multi-layered approach will reveal how children's literature functions as both a mirror and a mechanism for processing cultural disasters. This project contributes methodologically by developing age-appropriate content analysis techniques that can be applied across multilingual children's literature corpora. The research will develop new computational tools for analysing pedagogical discourse in digital humanities contexts, while establishing protocols for the ethical analysis of texts designed for vulnerable populations. This multi-layered approach will reveal how children's literature functions as both a mirror and a mechanism for processing cultural disasters.

This project contributes methodologically by developing age-appropriate content analysis techniques that can be applied across multilingual children's literature corpora. The research will develop new computational tools for analysing pedagogical discourse in digital humanities contexts, while establishing protocols for the ethical analysis of texts designed for vulnerable populations.

Expected Results

This research expects to uncover Nordic-Baltic children's literature likely contains culturally specific disaster motifs reflecting regional environmental conditions (harsh winters, maritime dangers) and historical traumas (occupation, language suppression). Expected findings include heightened emphasis on community resilience in Baltic texts reflecting occupation experiences, and environmental disaster motifs in Nordic texts reflecting climate extremes; Analysis will likely reveal temporal spikes in specific disaster themes corresponding to historical events (World War II occupation, Cold War tensions, environmental crises); Post-disaster periods are expected to show increased pedagogical emphasis on preparedness and resilience in children's materials. Translated texts are anticipated to reveal both adaptation strategies (localization of disaster contexts) and persistent universal elements (archetypal survival narratives). Border regions may show hybrid disaster narratives blending multiple cultural traditions; Comparative analysis across languages will likely identify culturally specific euphemisms, metaphors, and narrative techniques for communicating catastrophe to children while maintaining psychological safety; Diachronic analysis is expected to reveal shifts from religious and moral frameworks (medieval and early modern periods) toward psychological and scientific frameworks (20th-21st centuries) in children's disaster narratives; Theoretically, the project expands disaster studies into children's literature scholarship, revealing how societies utilise juvenile texts as both a protective buffer and a transmission vehicle for catastrophic knowledge. By demonstrating how abundance in digitised materials can uncover previously invisible patterns of cultural disaster processing, this research illustrates how children's literature contributes to cultural reproduction and trauma management.

The findings will illuminate the hidden pedagogical functions of disaster narratives in cultural transmission, revealing how Nordic and Baltic societies have used children's literature to maintain cultural continuity during periods of catastrophic disruption. This knowledge has contemporary relevance for understanding how digital cultures might better prepare young audiences for climate change and other emerging global disasters.

By mining fragmented digital collections of overlooked children's texts, this research demonstrates how abundance in digitised materials can reveal previously invisible patterns of cultural disaster processing. The project ultimately uncovers how societies encode disaster wisdom within children's literature, showing how these seemingly simple texts serve complex cultural functions in preparing communities for catastrophic futures while preserving essential cultural knowledge across generational boundaries. This investigation of non-canonical children's literature thus offers new insights into both digital humanities methodologies and artistic approaches to disaster resilience in Nordic and Baltic contexts.

References

1.Alexander, J. C. (2012). Trauma: A social theory. Polity Press.

2.Bettelheim, B. (1976). The uses of enchantment: The meaning and importance of fairy tales. Vintage Books.

3.Buell, L. (2001). Writing for an endangered world: Literature, culture, and environment in the U.S. and beyond. Belknap Press.

4.Gulliksen, H. (2015). The National Library of Norway: From analogue to digital. Alexandria: The Journal of National and International Library and Information Issues, 26(2), 119–130.

5.Jockers, M. L. (2013). Macroanalysis: Digital methods and literary history. University of Illinois Press.

6.Kangas, S., & Pasanen, T. (2018). Mining the treasures of the past: Digital archives and data processing in Finnish cultural heritage institutions. Journal of Cultural Heritage, 34, 247–253.

7.LaCapra, D. (2001). Writing history, writing trauma. Johns Hopkins University Press.

8.Moretti, F. (2005). Graphs, maps, trees: Abstract models for literary history. Verso.

9.Nikolajeva, M. (2018). Aesthetics and Ideology in Children's Literature. Routledge.

10.Stephens, J. (1992). Language and ideology in children's fiction. Longman.

11.Underwood, T. (2019). Distant horizons: Digital evidence and literary change. The University of Chicago Press.

12.Zipes, J. (2012). Relentless progress: The reconfiguration of children's literature, fairy tales, and storytelling. Routledge.

Session 2D — 10:30–12:30

10:30–10:50 **Mapping Diversity in Norwegian Literature for Children and Young Adults (2000–2025)**
Lars G Bagøien Johnsen, Kristin Ørjasæter

10:50–11:10 **Who is Sylvia? A comparison of plays by Albee and Rattigan**
Maria Bekker-Nielsen Dunbar, Manex Agirrezabal Zabaleta

11:10–11:30 **Molecules of a story: Community detection in narrative networks to unearth micro-narratives**
Kasper Fyhn, Rebekah Baglini

11:30–12:00 **Using LLMs to uncover hidden patterns in the contestation of religious authorities across a corpus of medieval inquisition records, 1243–1522**
David Zbíral, Zoltan Brys, Robert L. J. Shaw, Gideon Kotzé

12:00–12:30 **Trump, Catastrophe, and the Millennium: A Data-driven Study of Eschatological Discourses Surrounding Trump on 4chan's /pol/ Board**
Lauritz Holm Petersen

10:30–10:50 *SHORT PAPER* **[45]**

Mapping Diversity in Norwegian Literature for Children and Young Adults (2000–2025)

Lars G Bagøien Johnsen[1], Kristin Ørjasæter[2]

[1] *National Library of Norway, Norway*

[2] *Norwegian Institute for Children's Books*

Keywords: *diversity, children's literature, LLMs, digital methods*

Purpose

This paper examines how cultural-policy ambitions for diversity correspond to identity representation in Norwegian children's and young adult fiction published between 2000 and 2025.

We compare two corpora. The first is a large national corpus of Norwegian-language fiction and non-fiction for children and young people digitized by the National Library of Norway and made available through NB DH-LAB. The corpus supports frequency lists, concordance extraction, and collocation analysis (Birkenes & Johnsen 2025; Johnsen 2019).

The second is a smaller policy corpus consisting of two white papers from the Ministry of Culture: White Paper no. 10 (2011–2012), *Culture, Inclusion, and Participation*, and White Paper no. 18 (2020–2021), *Experience, Create, Share. Art and Culture for, with, and by Children and Youth.*

The policy corpus provides a normative framework for defining diversity. The fiction corpus provides empirical narrative practice. The study asks not only whether identities appear, but how they function contextually.

Method: From Markers to Luminons

Earlier analyses relied primarily on marker frequency and collocation. The present study extends this approach by introducing structured fragment-level annotation.

We use the term *luminon* to denote a qualified concordance fragment—a locally bounded textual configuration in which an identity marker is interpreted according to a predefined annotation schema. In methodological terms, luminons are instances of structured LLM-assisted annotation (cf. Mimno 2026).

Seed terms were drawn from demographic statistics, pedagogy, and policy language. These were expanded through collocation mining within the DH-LAB environment (Birkenes & Johnsen 2025). Concordances were then harvested around each marker.

Instead of counting lexical occurrences alone, we classify each concordance fragment according to interpretive distinctions such as:

literal vs. metaphorical

practice vs. theme

identity vs. event

backgrounded norm vs. marked identity

For example:

“Han går i synagogen” → religion as lived practice

“Burgerbutikken ligger bak synagogen” → religion as spatial reference

“Han var blind av raseri” → metaphorical

“Hun er blind og bruker stokk” → denotative disability

Few-shot calibrated prompts assist classification, and ambiguous cases trigger human review.

For each book, statistics are generated from luminons rather than raw tokens. We compute:

Presence: whether at least one qualified luminon of a given type occurs in a book

Intensity: the number of qualified luminons relative to text length

This allows quantitative comparison while filtering out metaphorical or incidental uses. Optional Louvain clustering visualizes shared qualified identities across books.

Our approach aligns with recent large-scale annotation efforts in literary studies, where LLMs are used to scale structured interpretation (Mimno 2026). However, luminons are anchored in concordance extraction and schema-driven labeling, ensuring bounded interpretation and reproducibility.

Cultural Policy as Framework

The white papers were analyzed using the same concordance workflow. We extracted definitional passages concerning “diversity,” “children,” and “youth,” and mapped how literature is positioned within policy discourse.

In 2011, literature is framed primarily as a tool for inclusion and democratization, especially for immigrant youth. By 2021, children are positioned more explicitly as artistic actors, and diversity becomes embedded within institutional structures.

These policy formulations informed the operational dimensions applied in the fiction corpus.

FindingsCountries and Languages

Representation mirrors translation streams. Anglophone and Nordic contexts dominate. Large immigrant groups in Norway, such as Polish backgrounds, are weakly represented.

Pakistan stands out: when present, it is narratively central and frequently associated with transnational mobility and identity negotiation. Here, migration and religion luminons cluster.

Religion

Religion appears infrequently overall. Protestant Christianity typically functions as backgrounded seasonal or habitual practice. Minority religions more often appear as thematic framing (e.g., Holocaust narratives; migration contexts).

The asymmetry is functional: majority religion appears normalized, minority religion marked.

Sexuality and Gender Identity

Queer identities are explicitly labeled and narratively foregrounded, particularly in YA developmental arcs. Heterosexuality is largely inferred through unmarked romantic plots. The asymmetry lies in markedness: heterosexuality functions as default norm.

Disability and Deafness

Assistive-device markers are more reliable than base terms such as “blind” or “deaf,” which frequently occur metaphorically. Raw lexical counts therefore overestimate representation. Luminon qualification reveals comparatively sparse denotative disability.

Limitations

Certain identities (e.g., Sámi) require entity linking rather than naïve lexical markers due to linguistic ambiguity.

Illustrated books present a methodological limitation: diversity is often iconographic rather than lexical, and thus undercounted in a word-based pipeline.

Presence and intensity do not capture stance or stereotyping. A second-stage luminon schema focusing on agency and evaluative framing is under development.

Contribution

Methodologically, this study extends earlier corpus-based literary analysis within the National Library’s DH infrastructure (Johnsen 2019; Birkenes & Johnsen 2025) by integrating structured LLM-assisted fragment annotation.

Substantively, it shows that diversity in Norwegian children’s and YA fiction is present but unevenly structured. Majority identities function as background norms, while minority identities are more often explicitly marked and thematically framed. Intersectionality within single titles remains limited.

The gap between policy ambition and literary practice is patterned rather than absolute.

References

Birkenes, M. B., & Johnsen, L. G. (2025). Corpus and the Bibliography: NB DH-LAB as an Infrastructure for Text and Metadata I J.-M. Hanssen & S. Furuseth (Red.), The Hermeneutics of Bibliographic Data and Cultural Metadata. Oslo: Notabene.

Bishop, R.S. (1990). Mirrors, Windows and Sliding Glass Doors. I Perspectives. Choosing and Using Books for the Classroom 6 (3). https://scenicregional.org/wp-content/uploads/2017/08/Mirrors-Windows-and-Sliding-Glass-Doors.pdf [Lesedato 11.02.2020]

Johnsen, L. (2019). «Eldre bøker i den digitale samlingen. Et elektronisk blikk på tekster fra perioden 1650-1850». I Litterære verdensborgere. Transnasjonale perspektiver på norsk bokhistorie 1519-1850. A.M.B. Bjørkøy, R. Hemstad, A. Nøding & A.B. Rønning (Eds.). Oslo: Nasjonalbiblioteket, 2019, 190-214.

Mimno, David. 2026. Crossing the Room to Crossing the World: Character Movement Annotation at Scale. Talk presented at TEXT: Center for Contemporary Cultures of Text, Aarhus University, 28 January 2026.

White papers

Meld. St. 10 (2011-2012) Kultur, inkludering og deltaking.

Meld. St. 18 (2020-2021) Oppleve, skape, dele. Kunst og kultur for, med og av barn og unge.

10:50–11:10 *SHORT PAPER* **[46]**

Who is Sylvia? A comparison of plays by Albee and Rattigan

Maria Bekker-Nielsen Dunbar[1], Manex Agirrezabal Zabaleta[2]

[1] *University of Heidelberg, Germany*

[2] *University of Copenhagen, Denmark*

Keywords: *quantitative drama analysis, stylometry, network analysis*

We analyse two plays named "who is Sylvia?". One is written by Terrance Rattigan and is set in England and one is written by Edward Albee and set in the United States. The goal of this work is to analyse these two plays and determine the characteristics of the plays and the character Sylvia. Both plays are about a protagonists engaging in sexual acts outside of their marriage with the titular character functioning as the paramore. Both are named for a poem in a Shakespeare play where Sylvia starts out as a speaking character but becomes non-speaking. In both our plays, Sylvia is a non-speaking character and we wish to understand the characteristics of this character. First we determine how similar the two plays are through stylometric analysis, comparing the writing of Albee and Rattigan. This is to gain an understanding of whether the temporal distance and variance of English has an impact. Second, we perform a character presence analysis considering how often each character takes a turn to speak in each act. This is to gain an understanding of secondary characters in relation to the protagnist. Last, but not least, we investigate the network of characters to determine the importance of the character of Sylvia. This is to understand the importance of her role, thought it is non-speaking. All together, this provides a first-order approach in understanding the characteristics of Sylvia and so answering the titular question. Both plays have a tripartite structure, meaning they consist of three acts. While the structure is the same, the length is not; the Rattigan piece is almost twice as long. Albee's play is almost entirely dialogue while Rattigan's play contains more scene exposition. We use the stylometric analysis approach of Mendenhall (1887) and plotted the number of letters in a word used against the number of words with that amount of letter (frequency) for a sample of 800 words. We find from this simple approach that the two plays are similar: the authors' writing styles have a similar shape and the three acts are similar also. We found this to be a somewhat surprising result as the plays are of different lengths in addition to being from different decades and written in different English language variations. We want to compare this with television show episodes with the same name ("Who is Sylvia?") and use other stylometric methods. We labelled the speaking roles by order following the Bourhis et al. (2024) method and analysed this labelled string. We observe that the protagonists (Martin in Albee and Mark in Rattigan) speak the most lines of dialogue. In Albee's play the best friend role (Ross, whose equivalent in Rattigan is called Oscar) follows a U-shape, appearing less in the middle of the play, while the son's (Billy) presence increases by act. He is mentioned in the first act and appears in a speaking role in later acts. Comparing the spouse role, Stevie plays a persistent role in Albee's play while Rattigan's Caroline only speaks in the third act. In Rattigan's play we need to collapse certain characters to understand the presence of their role and compare with Albee. There are two sets of love interests: for the protagonist this is Daphne, Nora, and Doris and for Oscar this is Ethel, Bubbles, and Chloe. The first set speaks most initially, while the second set speaks more as the play progresses. Our preliminary results suggest that the plays differ in speech and so we identified a characteristic that is different between the two. We will analyse not just the amount of times speaking but also the duration of speech in future work to determine whether this affects our results. For each line of dialogue, we extracted their mentions of other characters by name. We created networks of characters using these mentions to link characters together. We visualise this as directed graphs. The network is not a co-presence network

since Sylvia does not appear in the play but rather a "who mentions whom" network. This approach is inspired by the DraCor project (Trilcke et al. 2015). The network decreases in size for Albee and increases for Rattigan. We calculate centrality measures to determine the importance of Sylvia summarised as a value that can be compared. Based on these measures, she takes a particularly important role in Rattigan's second act. The centrality scores seem to be similar for Albee's first and second acts, with a difference seen in the last act, where only one edge links to Sylvia. Based on eigenvector centrality, Sylvia seems to be most connected to other nodes in the first act of Albee and the second act of Rattigan, which corresponds to the number of mentions. The degree centrality represents the number of direct connections, and confirms Sylvia having great importance in the second act for Rattigan and more of a constant connection in Albee. Overall, we find that the plays are thematically similar and seem to be structured similarly. However, there are differences in dialogue and connections between characters that suggest who Sylvia is and how she drives plot is not the same for the two plays.

References

Mendenhall, T. C. 1887. "The Characteristic Curves of Composition." Science ns-9 (214s): 237–246. https://doi.org/10.1126/science.ns-9.214S.23

Bourhis, P., A. Boussidan, C. Fournial, and P. Gambette. 2024. "Detecting semantic or structural similarities for theater play comparison."

Trilcke, P., Fischer, F., and Kampkaspar, D. 2015 "Digital Network Analysis of Dramatic Texts" https://doi.org/10.5281/zenodo.3627710

11:10–11:30 *SHORT PAPER* **[47]**

Molecules of a story: Community detection in narrative networks to unearth micro-narratives

Kasper Fyhn[1,2,3], Rebekah Baglini[1,2]

[1] *Department of Linguistics and Cognitive Science, Aarhus University, Denmark*
[2] *TEXT - Center for Contemporary Cultures of Text, Aarhus University, Denmark*
[3] *Center for Humanities Computing, Aarhus University, Denmark*

Keywords: *narrative graph, rare entities, community detection, natural language processing*

Automatically extracted narrative networks – graphs with entities as nodes and their relationships as edges – can reveal structures of a narrative through salient entities and how they relate (Labatut and Bost 2019; Tangherlini et al. 2020; Holur et al. 2021). In such an entity-centric approach (as opposed to event-centric), the structure of the network reflects the narrative in terms of associations and events involving the extracted entities. Such networks naturally foreground central narrative structures — salient entities and their strong relations. But a narrative is more than the central structures that everything else revolves around, builds upon and adds to.

This work is concerned with the *everything else* – brief sub-plots, small clusters of descriptions, or associations between minor characters that make a fictional world come alive, but which may go under the radar in computational studies of text. We refer to these as *micro-narratives* – delimited peripheral narrative structures, localized configurations of entities that together constitute recognizable subplots or thematic clusters, identified as communities of rare entities in the narrative network. We present an approach to unearth such micro-narratives.

The micro-narratives involve rare entities with limited textual presence. In network-based approaches to narrative, entity salience is typically determined by frequency of occurrence or network centrality measures, which naturally foregrounds dominant entities and the overarching narrative structures they constitute. But rarer entities are not only overshadowed by dominant ones — they are also lost among each other in the long tail of many but rare entities (Baayen 2001). Identifying meaningful peripheral structures in this abundance is the core challenge this work addresses. After all, as molecules of a larger narrative structure, they help give it depth and weight.

We create a narrative graph from a text collection by performing Named Entity Recognition extended by Noun Phrase Chunking with SpaCy (Honnibal et al. 2020) to extract everything from salient to inconspicuous entities in a broad sense — characters, locations, objects, and concepts — and taking

these as nodes in the graph. Since we are interested in rare entities, we employ little filtering and are hesitant to resolve different mentions of potentially identical entities too aggressively, so as to not distort the weak signals. Edges between the entity nodes are created by counting co-occurrences between extracted entities within some defined window of text. Since we here favor the rarities, we weight the edges using pointwise mutual information (Church and Hanks 1990) – a measure of association, the primary weakness of which is its inflation for rare events. For this purpose, its weakness becomes its strength. The result is a weighted, undirected graph with heterogeneous entity types as nodes, and PMI-weighted co-occurrence as edges.

Given a graph that favors weak relational signals, we employ community detection techniques (both with and without overlapping communities; Palla et al. 2005; Blondel et al. 2008) to identify tightly connected sections of the graph as structural traces of underlying micro-narratives. Still, irrelevant or erroneously extracted entities can dominate in the extracted communities, as we search for meaningful structures among noise in the long tail. To separate the wheat from the chaff, a range of structural measures aid in their identification, including their textual confinement and their connectivity to salient nodes in the broader narrative.

The method is demonstrated on *The Lord of the Rings* as a literary work with rich narrative complexity. We further apply it to a corpus of Reddit comments about ChatGPT in the months after its launch to explore the extent to which the method generalizes beyond fictional narrative to emergent public discourse. We evaluate the method qualitatively by examining extracted micro-narratives and their characteristics across the two registers and with two different community detection algorithms. The two cases highlight structural similarities and differences as well as strengths and weaknesses of the method.

The proposed method offers a conceptually simple yet effective way of targeting peripheral narrative structures, using well-known techniques in a novel creative way. In doing so, it contributes to bridging distant and close reading by surfacing the kind of fine-grained narrative details that can be lost in large-scale computational text analysis.

References

Baayen, R. Harald. 2001. Word Frequency Distributions. Vol. 18, edited by Nancy Ide and Jean Véronis. Text, Speech and Language Technology. Springer Netherlands. https://doi.org/10.1007/978-94-010-0844-0.

Blondel, Vincent D., Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. 2008. “Fast Unfolding of Communities in Large Networks.” Journal of Statistical Mechanics: Theory and Experiment 2008 (10): P10008. https://doi.org/10.1088/1742-5468/2008/10/P10008.

Church, Kenneth Ward, and Patrick Hanks. 1990. “Word Association Norms, Mutual Information, and Lexicography.” Computational Linguistics 16 (1): 22–29.

Holur, Pavan, Shadi Shahsavari, Ehsan Ebrahimzadeh, Timothy R. Tangherlini, and Vwani Roychowdhury. 2021. “Modelling Social Readers: Novel Tools for Addressing Reception from Online Book Reviews.” Royal Society Open Science 8 (12): 210797. https://doi.org/10.1098/rsos.210797.

Honnibal, Matthew, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-Strength Natural Language Processing in Python. https://doi.org/10.5281/zenodo.1212303.

Labatut, Vincent, and Xavier Bost. 2019. “Extraction and Analysis of Fictional Character Networks: A Survey.” ACM Comput. Surv. 52 (5): 89:1-89:40. https://doi.org/10.1145/3344548.

Palla, Gergely, Imre Derényi, Illés Farkas, and Tamás Vicsek. 2005. “Uncovering the Overlapping Community Structure of Complex Networks in Nature and Society.” Nature 435 (7043): 814–18. https://doi.org/10.1038/nature03607.

Tangherlini, Timothy, Shadi Shahsavari, Behnam Shahbazi, Ehsan Ebrahimzadeh, and Vwani Roychowdhury. 2020. “An Automated Pipeline for the Discovery of Conspiracy and Conspiracy Theory Narrative Frameworks: Bridgegate, Pizzagate and Storytelling on the Web.” PLOS ONE 15 (6): e0233879. https://doi.org/10.1371/journal.pone.0233879.

11:30–12:00 *LONG PAPER* **[48]**

Using LLMs to uncover hidden patterns in the contestation of religious authorities across a corpus of medieval inquisition records, 1243–1522

David Zbíral, Zoltan Brys, Robert L. J. Shaw, Gideon Kotzé

Masaryk University, Czech Republic

Keywords: *religious authority contestation, inquisition records, information extraction, Large Language Models*

Introduction

The direct contestation of religious authorities was not a dominant topic of medieval inquisition records, contrary to both common expectations and the usual extent of the notion of heresy, which has, from 1075 onwards, also encompassed any sustained disobedience to ecclesiastical authority. Most of the medieval inquisitorial documentation in fact focuses on other transgressions, such as illicit meetings, rituals, doctrinal heterodoxy, or indirect forms of challenging ecclesiastical authority and its salvific power, such as the preference of dissident ministers over those affiliated to the church, and the criticism of sacraments (Arnold 2001: 152; Biller 2001: 314; Sackville 2011: 118–9). Nevertheless, heresy trial records contain valuable information on direct and explicit contestation of the authority of the church, papacy, religious orders, and priests; only it is overshadowed by other topics and dispersed across a large body of evidence. The volume of this textual material certainly exceeds the possibilities of unassisted close reading. In addition, the diversity of expressions of religious authority contestation defies traditional corpus query methods, especially if we are not interested in the mere presence or absence of the topic itself (e.g., in inquisitorial questions about it, or in reports about other – potentially repeating – suspects), but in whether deponents themselves confessed to have held or to had previously held such opinions. Recent methodological work shows that large language models (LLMs) can support the systematic analysis of large quantities of historical textual data by automating core analytical tasks such as classification and information extraction. This development enables researchers to address well-specified research questions at scale (Barros et al. 2025). Applied studies of LLM integration demonstrate how LLMs can streamline common analytical steps such as document classification and extraction of relevant content (Fischer and Biemann 2024). At the same time, this literature also emphasizes that methodological rigour requires careful specification of prompts and evaluation metrics (Stewart and Sinha 2025).

Knowledge gap

There is a general awareness that the contestation of religious authorities is part and parcel of different dissident milieus (Piron 2003; Taviani-Carozzi 2003), but there seems to be comparatively little interest in its broader patterns. True, scholarship has certainly displayed some interest in the forms of such contestations in inquisition records (Albaret 2003) and even in local developments of anticlerical attitudes over time (Biget 2003). However, to the best of our knowledge, no attempt has previously been made to (1) quantify how frequently such verbal challenges to religious authorities occur in a larger body of inquisitorial evidence and (2) examine factors that correlate with their presence or absence. Regarding the latter, some of the most relevant factors that might realistically play a role are:

Religious culture. We might reasonably expect that different milieus would display variation in the frequency with which defendants affiliated to them would confess the contestation of religious authorities.

Predominantly urban milieu. We might assume that urban political struggles and political cultures could have supported the incidence of criticism of religious authorities.

Later period. With the development of the public sphere in the Later Middle Ages (Guenée 2002; Hobbins 2009; Giraudet 2022; Barucci 2024), we might expect increased frequency of authority contestation over time.

Gender. We might assume that women would have had more reservations in expressing direct religious authority contestations, for example due to their more vulnerable social status.

In this study, we therefore examine whether any of these four factors is associated with the frequency of religious authority contestations.

Materials and methods

To constitute the corpus, we digitised modern source editions on a professional robotic scanner, optically recognized them with ABBYY FineReader software, constituted the main text by manually removing editorial material (footnotes, critical apparatus, etc.), segmented registers into individual documents, browsed and searched the texts for the most typical OCR errors, and transformed them into the plain-text format. In the case of two inquisition registers, we used the available digital-born editions.

For the acquisition of variable values, we used – depending on the variable in question – a combination of (1) human-assigned metadata and (2) zero-shot classification by Anthropic's Claude Sonnet 4 large language model (claude-sonnet-4-20250514). LLM-classifications were validated against independent human coding on random samples of 200 testimonies per variable, using identical coding rules. Agreement reached 85% for testimony identification, 94% for authority contestation, 99% for gender, and 84% for religious culture.

Overall, we analyzed 4,357 individual documents from 20 inquisition registers, spanning the period from 1243 to 1522 and covering various regions of Western Christendom (namely, today's South-Western France, North-Central Italy, Switzerland, and England, our choices being partly predetermined by the provenance of most of the documentation extant to this day).

For data analysis, we used regression analyses at both register and testimony levels, and complemented them with random forest variable importance estimation. To account for some of the possible confounding factors, we also included as control variables: (1) text length, (2) Measure of Textual Lexical Diversity – MTLD (McCarthy and Jarvis 2010), and – as at least a basic proxy for inquisitorial pressurizing by questions – (3) the number of recorded inquisitorial questions, proportioned towards the typical number of questions per testimony in the given register.

Results and discussion

Testimonies associated with the traditionally demarcated dissident religious cultures of the Beguins (AME [Average Marginal Effect] = 0.41), Lollards (AME = 0.26), other heterodoxy (AME = 0.12), and Waldensians (AME = 0.10) were more likely to contain contestations of religious authorities than those from non-heterodox deponents, while those associated with Cathars, Apostles and Guglielmites showed no significant effect. We interpret this result by proposing a typology, which divides this traditional typology of dissident religious milieus more clearly into (a) reformistic, which seem to have been more concerned with the state of the church, and thus engaging more often in its criticism, and (b) separatistic, which seem to have felt less of such an urge (relatively speaking).

Registers predominantly documenting urban centers were less likely (AME = -0.09) to contain religious authority contestations. This is a surprising result: we expected a positive association. Thus, within our corpus, it is registers with predominantly rural settings which tend to contain more of religious authority contestation.

Furthermore, as expected but so far never demonstrated using quantitative methods, registers from a later period were more likely (AME = 0.08) to contain the contestation of religious authorities, thus showing that this topic somewhat grew in importance. We relate this result to the expanding importance of the public sphere and laypeople's voice as time progressed towards the Reformation.

Lastly, the defendants' gender did not display any effect: female defendants were not less likely to formulate religious authority contestation than male defendants.

Generally, we hope to have illustrated the potential of LLM-based extraction of sparse information from larger textual corpora to answer research questions of interest in cultural history, in this case, in the study of large-scale patterns in historical forms of resistance. This paper thus constitutes a broader contribution to the discussion about the opportunities offered by LLMs, which, for some tasks, bears comparison with the quality of human annotation (Karjus 2025). If used with clear definitions of variables in prompts and complemented by rigorous human cross-validation, they can provide data to answer some of the broader questions in cultural history that we would hardly be able to answer without such assistance.

References

Albaret, Laurent (2003) 'L'anticléricalisme dans les registres d'inquisition de Toulouse et de Carcassonne au début du XIVe siècle', in L'anticléricalisme en France méridionale, Cahiers de Fanjeaux, 38 (Toulouse: Privat), 447–70.

Arnold, John H. (2001) Inquisition and Power: Catharism and the Confessing Subject in Medieval Languedoc, The Middle Ages (Philadelphia: University of Pennsylvania Press).

Barros, Cauã Ferreira, et al. (2025) 'Large Language Model for Qualitative Research -- A Systematic Mapping Study', arXiv:2411.14473, preprint, arXiv, 6 March, https://doi.org/10.48550/arXiv.2411.14473.

Barucci, Teresa (2024) 'The Medieval Public Sphere and the Response to a Condemnation for Heresy in Bologna, 1299', The English Historical Review, 139/598–599: 651–79, https://doi.org/10.1093/ehr/ceae109.

Biget, Jean-Louis (2003) 'L'anticléricalisme des hérétiques d'après les textes polémiques', in L'anticléricalisme en France méridionale, Cahiers de Fanjeaux, 38 (Toulouse: Privat), 405–45.

Biller, Peter (2001) 'Through a Glass Darkly: Seeing Medieval Heresy', in Peter Linehan and Janet L. Nelson, eds, The Medieval World (London; New York: Routledge), 308–26, https://doi.org/10.4324/9781315102511-21.

Fischer, Tim, and Chris Biemann (2024) 'Exploring Large Language Models for Qualitative Data Analysis', Proceedings of the 4th International Conference on Natural Language Processing for Digital Humanities 423–37, https://doi.org/10.18653/v1/2024.nlp4dh-1.41.

Giraudet, Luke (2022) Public Opinion and Political Contest in Late Medieval Paris: The Parisian Bourgeois and His Community, 1400-50 (Turnhout, Belgium: Brepols Publishers n.v.).

Guenée, Bernard (2002) L'opinion Publique à La Fin Du Moyen Âge. D'après La Chronique de Charles VI Du Religieux de Saint-Denis, Pour l'histoire (Paris).

Hobbins, Daniel (2009) Authorship and Publicity Before Print: Jean Gerson and the Transformation of Late Medieval Learning (n.p.: University of Pennsylvania Press).

Karjus, Andres (2025) 'Machine-Assisted Quantitizing Designs: Augmenting Humanities and Social Sciences with Artificial Intelligence', Humanities and Social Sciences Communications, 12/1: 277, https://doi.org/10.1057/s41599-025-04503-w.

McCarthy, Philip M., and Scott Jarvis (2010) 'MTLD, Vocd-D, and HD-D: A Validation Study of Sophisticated Approaches to Lexical Diversity Assessment', Behavior Research Methods, 42/2: 381–92, https://doi.org/10.3758/BRM.42.2.381.

Piron, Sylvain (2003) 'La critique de l'Église chez les Spirituels languedociens', in L'anticléricalisme en France méridionale, Cahiers de Fanjeaux, 38 (Toulouse: Privat), 77–109.

Sackville, Lucy J. (2011) Heresy and Heretics in the Thirteenth Century : The Textual Representations (Woodbridge; Rochester: York Medieval Press).

Stewart, Spencer Dean, and Sanskriti Sinha (2025) 'Retrieving Information from Unstructured Historical Sources Using Large Language Models', Computational Humanities Research, 1: e17, https://doi.org/10.1017/chr.2025.10019.

Taviani-Carozzi, Huguette (2003) 'Anticléricalisme et ecclésiologie: Pierre le Vénérable, le moine Guillaume et les hérétiques', in L'anticléricalisme en France méridionale, Cahiers de Fanjeaux, 38 (Toulouse: Privat), 329–53.

12:00–12:30 *LONG PAPER* **[49]**

Trump, Catastrophe, and the Millennium: A Data-driven Study of Eschatological Discourses Surrounding Trump on 4chan's /pol/ Board

Lauritz Holm Petersen[1,2]

[1] *Aarhus University*

[2] *University of Southern Denmark, Denmark*

Keywords: *Trump, Eschatology, 4chan /pol/, semantic network, HMM discursive regime analysis*

In this paper, I present a data-driven examination of eschatological discourses surrounding Donald J. Trump within the subcultural far-right milieu of 4chan's "Politically Incorrect" (/pol/) board. While frequently framed as an eschatologically significant entity in line with biblical premillennialist traditions, Trump's signature mix of deal-making geopolitics and seemingly improvised populist rhetoric has also motivated secular-catastrophic framings that position him as a catalyst for societal collapse and global disaster. Drawing on a dataset of 242,000 eschatologically relevant posts spanning 2014–2023, the study employs a mixed-methods approach combining TF-IDF-based co-occurrence network analysis, Hidden Markov Modeling, topic modeling, and qualitative close readings to investigate the content and longitudinal dynamics of dominant eschatological discourses surrounding Trump on the subcultural fringe. The analysis confirms the existence of two lexically distinct but interacting eschatological frames, biblical premillennialism and secular catastrophism, and shows that Trump, as a central eschatological figure, oscillates between these frames in response to key socio-political events, including the 2016 and 2020 elections, U.S. interventions in Syria, and the COVID-19 pandemic. In contrast to dominant portrayals, these findings illuminate Trump's ambivalent position as a political figure within the subcultural far right and illustrate how /pol/ operates as a generative space for negotiating and adapting eschatological frames and oppositional identities in continuous dialogue with shifts in the broader socio-political milieu.

Background

In a Western context, eschatological narratives have long functioned as widely distributed interpretive frameworks through which communities make sense of socio-political events (Barkun, 2013, p. 16; Harding, 1994; Wojcik, 1997, p. 37). Scholars across disciplines have emphasized that such narratives are remarkably flexible constructs, readily adapted to reflect unfolding events and shifting circumstances (Harding, 1994, p. 35; Landes, 2011; Worsley, 1970 [1957], p. 294). A clear manifestation of this flexibility is the persistent tendency to frame religious and political leaders in eschatological terms. From anti-Catholic portrayals of the pope as the Antichrist to depictions of American figures such as Ronald Reagan, Barack Obama, and, most recently, Donald Trump, as either divinely ordained saviors or deceitful agents of evil, eschatological narratives are easily adapted to contemporary matters. (Bond & Neville-Shepard, 2023; Jenkins, 2003; Whisker, 2012).

A substantial body of scholarship (Balmer, 1988; Boyer, 1994; O'Leary, 1994; Stewart & Harding, 1999; Wojcik, 1996, 1997) identifies two major strands of eschatological thought: Christian premillennialism and secular catastrophism, both of which became rearticulated during the rise of Donald Trump. Trump's rhetoric regarding "hordes" of immigrants, the rampant rise of Islam and communism, a corrupted youth, and a post-apocalyptic hellscape of "rusted-out factories" mapped seamlessly onto premillennialist interpretations of the geopolitical landscape, in which the rise of evil and societal collapse signals the approaching Millennium and Christ's Second Coming. But scholarly accounts of Trump's political success also point beyond the traditional evangelical base to the broader far-right ecosystem, particularly the online subcultural far right, as a key factor. These communities are as well preoccupied with visions of impending collapse and transformation but draws on a broader repertoire of end time visions, from human enslavement to race wars and demographic decline (Petersen & Baun, 2025). Their eschatological orientation thus extends beyond premillennialist conceptions of the end times, encompassing ideas more akin to secular catastrophism (Stewart & Harding, 1999; Wojcik, 1997). Given the diversity of eschatological orientations and their inherent political ambivalence, the online subcultural far right offers a compelling context for examining how competing eschatological narratives are mobilized to frame contemporary political figures. Among this broader ecology of platforms, 4chan's /pol/ board stands out as arguably the most influential site hosting controversial political discussions and generating ideological content that has migrated beyond the online fringe to shape the American political landscape. The board's unique infrastructure of user anonymity and content ephemerality, combined with its stated purpose of discussing "news, world events, and political issues" (Anonymous, 2017), makes it an ideal milieu for alternative news reporting and the negotiation of oppositional ideologies (Bernstein et al., 2011; Lapidot-Lefler & Barak, 2012).

The availability of longitudinal user-generated data from platforms like 4chan presents a unique opportunity to examine, in a data-driven way, processes of ideological adaptation in response to unfolding socio-political events. This study asks: What role does Trump play in the eschatology of the online subcultural far right, and how is this narrative positioning shaped by real-world developments over time? Addressing these questions, the study aims to provide a more nuanced understanding of a key segment of Trump's electoral base, while contributing data-driven insights into the interpretive dynamics through which contemporary eschatology is adapted to shifting political contexts.

Methods

Data was collected from 4chan's /pol/ using a query-based strategy designed to isolate eschatologically relevant discussions. Using the 4CAT toolkit (Peeters et al., 2022), a seed list of common eschatological expressions was iteratively refined to include board-specific variants. This strategy emphasized event-centered expressions rather than doctrinal figures (e.g., Christ, Antichrist), allowing additional eschatological entities to emerge organically. The final curated list produced 242,596 posts spanning January 2014–April 2023.

Preprocessing involved standard text-cleaning operations tailored to 4chan text: lowercasing, punctuation and whitespace normalization, removal of reply markers, numeric strings, and stopwords (scikit-learn). To address 4chan's frequent content recycling, duplicates were removed, including near-duplicate posts detected using Locality-Sensitive Hashing and MinHash. Posts were sentence-segmented with spaCy, as /pol/ posts often span multiple semantic segments. Bigrams such as "end times" and "second coming" were merged into single tokens, and proper names were standardized (e.g., "Trump," "Donald," "Drumpf" → Donald_Trump). The final dataset comprised 950,410 sentences from 231,744 unique posts.

Analyses combined co-occurrence network analysis, Hidden Markov Modeling (HMM), and topic modeling.

- Network analysis: To obtain a static overview of central entities and relations in the eschatological discourse, co-occurrence networks were constructed with NetworkX from TF-IDF-weighted terms across the full dataset, as well as for four “transition periods” revealed by the trend analysis (Trump’s campaign 2015–2016, early presidency 2016–2017, COVID-19 onset 2020, and post-election turmoil 2020–2021). Networks were visualized in Gephi using ForceAtlas2, and Louvain modularity clustering identified semantic communities representing dominant frames.

- HMM trend analysis: To track temporal changes in Trump’s eschatological framing, monthly co-occurrence frequencies between the Trump entity and key lexical clusters were modeled with hmmlearn to detect latent “discursive regimes.” Model selection via AIC and BIC indicated five optimal states, representing shifts between strong/moderate premillennialist, secular-catastrophic, and ambivalent framings.

- Topic modeling and close readings: To explore the content and socio-political specifics of the four transition periods, BERTopic was applied. For each period, sentence embeddings were generated using a transformer-based model (all-MiniLM-L6-v2). Topics were manually aggregated into categories, and representative sentences and their threads were cross-referenced with archived /pol/ (4plebs) to retrieve explicitly referenced socio-political events or situations.

Findings

The analysis of eschatological discourse within the subcultural milieu of /pol/, and Trump’s positioning within it, yields several insights. First, both overall and period-specific co-occurrence networks reveal a systematic clustering of eschatological discourse into two lexically and thematically distinct frames: (1) a secular-catastrophic frame, oriented around large-scale societal and global disasters such as war, nuclear conflict, pandemics, and geopolitical instability; and (2) a biblical-premillennialist frame, focused on the Second Coming, the End Times, divine agents (e.g., God, Jesus Christ, the Antichrist), cosmological terms (e.g., heaven, hell), as well as Israel and Jews.

Second, the overall analysis indicates that Trump is generally a central figure in /pol/’s eschatological discourse. He appears not merely as a marginal or time-bound socio-political entity but as an integrated and persistent member of the eschatological network, positioned alongside narrative “core members” such as Antichrist, God, or Nuclear. Zooming in on Trump’s eschatological position during periods of discursive transition, the analyses indicate that his role oscillates between the two dominant clusters. Trump initially enters the eschatological discourse through his opposition to globalism, aligning loosely with the biblical-premillennialist frame. During his presidency, particularly amid interventionist foreign policies, he shifts more centrally into the secular-catastrophic frame, symbolizing an acceleration of societal collapse eagerly anticipated by segments of the /pol/ community. His eschatological significance diminishes during the early COVID-19 crisis, only to reemerge in relation to the 2020 election defeat, where his eschatological role is again evaluated within a premillennialist frame. These oscillations illustrate the subcultural far right’s ongoing and highly ambivalent efforts to reframe Trump’s significance in light of unfolding events and to adapt shared narratives of crisis and redemption to changing political realities. The findings complicate prevailing images of far-right spaces like /pol/ as uniformly pro-Trump milieus.

The analysis also draws attention to underexamined ways in which these eschatological traditions interact in practice. While lexically and to an extent ideologically and theologically distinct, the discursive clusters identified in the analysis frequently intersect, forming a general eschatological network of culturally salient items whose internal emphases shift with fluctuations in the socio-political milieu. In other words, biblical and secular discourses do not function as mutually exclusive silos but overlap in shared semantic fields and co-occur within the same pragmatic settings.

References

4plebs. 4plebs Archive. Accessed February 13, 2025. https://archive.4plebs.org/_/articles/credits/

Anonymous. “Welcome to /pol/ - Politically Incorrect.” 4chan. May 5, 2017. https://boards.4chan.org/pol/

Balmer, Randall. “Apocalypticism in America: The Argot of Premillennialism in Popular Culture.” Prospects 13 (1988): 417–433.

Barkun, Michael. A Culture of Conspiracy. University of California Press, 2013.

Bernstein, Michael S., Andrés Monroy-Hernández, Drew Harry, Paul André, Katrina Panovich, and Greg Vargas. “4chan and /b/: An Analysis of Anonymity and Ephemerality in a Large Online Community.” Proceedings of the International AAAI Conference on Web and Social Media 5, no. 1 (2011): 50–57.

Bond, Bradley E., and Ryan Neville-Shepard. “The Rise of Presidential Eschatology: Conspiracy Theories, Religion, and the January 6th Insurrection.” American Behavioral Scientist 67, no. 5 (2023): 681–696.

Boyer, Paul. When Time Shall Be No More: Prophecy Belief in Modern American Culture. Harvard University Press, 1994.

Harding, Susan. “Imagining the Last Days: The Politics of Apocalyptic Language.” Bulletin of the American Academy of Arts and Sciences 47, no. 3 (1994): 14–44.

Jenkins, Philip. The New Anti-Catholicism: The Last Acceptable Prejudice. Oxford University Press, 2003.

Landes, Richard. Heaven on Earth: The Varieties of the Millennial Experience. Oxford University Press, 2011.

Lapidot-Lefler, Noam, and Azy Barak. “Effects of Anonymity, Invisibility, and Lack of Eye-Contact on Toxic Online Disinhibition.” Computers in Human Behavior 28, no. 2 (2012): 434–443. https://doi.org/10.1016/j.chb.2011.10.014

O'Leary, Stephen D. Arguing the Apocalypse: A Theory of Millennial Rhetoric. Oxford University Press, 1994.

Peeters, Stijn, and Sal Hagen. “The 4CAT Capture and Analysis Toolkit: A Modular Tool for Transparent and Traceable Social Media Research.” Computational Communication Research 4, no. 2 (2022): 571–589.

Petersen, Lauritz Holm, and Phillip Stenmann Baun. 2025. “Clowns, Pills, & Boogaloos: A Data-Driven Study of Far-Right Eschatology on 4chan /Pol/.” Politics, Religion & Ideology 26 (2): 195–229. doi:10.1080/21567689.2025.2542770

Stewart, Kathleen, and Susan Harding. “Bad Endings: American Apocalypsis.” Annual Review of Anthropology 28, no. 1 (1999): 285–310.

Whisker, Daniel. “Apocalyptic Rhetoric on the American Religious Right: Quasi-Charisma and Anti-Charisma.” Max Weber Studies 12, no. 2 (2012): 159–184.

Wojcik, Daniel. “Embracing Doomsday: Faith, Fatalism, and Apocalyptic Beliefs in the Nuclear Age.” Western Folklore 55, no. 4 (1996): 297–330. https://doi.org/10.2307/1500138

Wojcik, Daniel. The End of the World as We Know It: Faith, Fatalism, and Apocalypse in America. NYU Press, 1997.

Worsley, Peter. The Trumpet Shall Sound: A Study of "Cargo" Cults in Melanesia. 1957. Reprint, Paladin, 1970.

Panel 2 & 3 — 13:30–15:30

13:30–14:30 **Finding a needle in a haystack – user experiences with digital heritage reuse**
Mart Alaru, Pille Runnel, Agnes Aljas, Pille Pruulmann-Vengerfeldt, Kai Pata, Natali Ponetajev

14:30–15:30 **Uncovering Hidden Infrastructures for Digital Humanities in GLAMs**
Mahendra Mahey, Pille Pruulmann-Vengerfeldt, Hans Dam Christensen, Berndt Clavier, Rikke Lie Halberg, Paula Bray

13:30–14:30 *PANEL* **[50]**

Finding a needle in a haystack – user experiences with digital heritage reuse

Mart Alaru[1], Pille Runnel[1], Agnes Aljas[1], Pille Pruulmann-Vengerfeldt[1,2], Kai Pata[3], Natali Ponetajev[4]

[1] *Estonian National Museum, Estonia*
[2] *Malmö University, Sweden*
[3] *Tallinn University, Estonia*
[4] *Estonian Literary Museum, Estonia*

Keywords: *Digital Cultural Heritage, Affordances, Data Infrastructures, Remix, Participation*

Panel: Finding a needle in a haystack – user experiences with digital heritage reuse Moderator: Mart Alaru (Estonian National Museum, Estonia) In this panel, we discuss the uses of digital heritage resources – user experiences, creative practices and outcomes emerging at the intersection of mediatised, archival and individual logics. The data for these presentations is derived from two interventions in Estonia: a digital heritage remix competition and a series of co-creative experiments with folk music groups and digitised folk music. The panel asks how do users create meaning for digital

heritage (collections) and how can it breach the gap between institutionally curated heritage and the user.

We look at how data infrastructures inform users in their creative endeavours, and, on the other hand, how user contexts and expectations contribute to the creative process and its outcomes. What frictions arise between creativity and digital heritage infrastructures? How can these infrastructures inspire and facilitate creative action and participation in general? From the users' different perspectives, what are the affordances that have led them to participate, from the level of the general digital landscape, e.g. social media, to the more specific digital heritage archives?

3 papers:

Needle in a Haystack – Experiences of Reusing Digital Heritage (Mart Alaru, Estonian National Museum, Estonia)

This paper analyses the use of digital heritage databases from the perspective of remix contestants whose task was to use the freely available digital heritage in Estonia for creative/artistic purposes. The panel contribution is based on 21 semi-structured interviews with contestants ranging from 20 to 67 years old. The idea of the contest was for users to create something new with the publicly available digital heritage databases (mostly digitised photos, videos and audio) on the topic of the 500th anniversary of the Estonian written word. The contest call was necessarily open in its expectations for the output, encouraging contestants to "go crazy", find and use digital heritage, and come up with anything remotely associated with the topic.

The paper brings out the user-centric view of heritage reuse, indicating the need for structure and meaning-making for digitized heritage that is predicated on the audience – if it were to be engaged with. Here, the organising element was the contest with its open-ended task, directing the users to look for something yet unspecified in the databases. Many interviewees reported that the contest gave them an excuse to do something creative, especially with it being open-ended. Still, without a clear idea of what to look for, many users found themselves lost in digital heritage databases, looking for a needle in a haystack.

The analysis combines the concept of affordances with mediatization, viewing the digital heritage remix contest and digital heritage databases as a collection of affordances that enabled the users to find an outlet for their creativity in a mediatised landscape that has affordances of its own – pushing and pulling attention to various interests. The paper is inspired by the idea that digital heritage is a space full of creative potential, able to spawn a wide variety of unique creations, giving heritage a new form and new life. This potential needs to be enabled by digital heritage databases, which implement goals and creative affordances of their own.

Youth Culture Meets Digital Cultural Heritage: A Content Analysis of Creative Reuse in Estonia (Agnes Aljas, Estonian National Museum, Estonia; Pille Runnel, Estonian National Museum, Estonia; Pille Pruulmann-Vengerfeldt, Estonian National Museum, Estonia and Malmö University, Sweden)

This paper discusses the transformation of digital heritage through its creative reuse, as it merges into and circulates within contemporary digital media. The creative remixing of the digital heritage resources involves multiple, simultaneous forms of change. The study draws on an experimental intervention at the Estonian National Museum, organised as an open competition that invited participants to creatively reuse digitised archival photographs, artworks, and other materials from the digitised collections of museums and archives. The competition resulted in 160 digital works, including memes, short videos, collages, and digital artworks, each reinterpreting heritage materials through the expressive grammars of contemporary digital formats. The paper presents qualitative analysis of the submitted creative works to take a closer look at the interpretative agency of participants when using digital heritage resources.

The transformation of heritage happening in artistic remixing has several interlinked dimensions. When digital archive materials are reborn as part of artworks and digital cultural formats, they undergo contextual, infrastructural, and epistemic shifts. Platform logics reshape heritage communication, turning heritage from document or evidence into aesthetic expression – from heritage as reference to heritage as expression. In this process, critical reflection often gives way to affective engagement, while heritage meanings become layered with nostalgia, irony, sincerity or parody.

Together, these processes generate hybrid artefacts in which historical meanings and contemporary expressions, digital affordances and personal creativity intersect. The circulation of heritage within

digital media illustrates how cultural memory is produced, mediated, and experienced. Through artistic remixes, contemporary digital formats act as active agents, shaping how heritage is perceived and reimagined in the present. The paper shows how digital heritage is not just preserved, it's performed, reinterpreted, and reimagined. It demonstrates how creative digital practices become a powerful form of cultural analysis, making it a valuable contribution to digital humanities scholarship concerned with digital culture, participatory media, and the evolving meanings of the past in networked spaces.

Collective Exploration of the Music Archive with Folk Musicians – a Case Study (Kai Pata, Tallinn University, Estonia and Natali Ponetajev, Estonian Literary Museum, Estonia)

Archives should enable transformative, deeply changing experiences, allowing new forms of cultural engagement, creative reuse, and community participation to emerge through the archive. On one hand, there is an expectation that community members could use the archives alone or together for shaping their own cultural identities. On the other hand, curated processes within the archive, which synthesise, mediate, and interpret folk music for communities, have been considered necessary.

We present an explorative case study of the cocreative music-making approach done using the resources of the Estonian Folklore Archive of Estonian Literary Museum (http://kirmus.ee) in collaboration with 78 folk instrument players of various ages and skill levels who attended the Võrumaa folk music camp. Estonian Folklore Archive stores folk music collected in the 19th-21st century that is accessible for individual search as well as through folklorist-moderated collections in the community hub of the archive. We experimented with 9 groups of folk musicians to see how these searchable music resources in the archive would inspire them to play and rearrange the music. Methodologically, we observed and video recorded the event and collected opinions of the process using the qualitative survey. Folk music learning has traditionally happened through the real-time meaning-making process in community settings, where situatedness in cultural practices and seeing different iterations of the same music being played were important. The music in archives has a temporal distance from modern-day players and may lack an easily accessible socio-cultural context, which hinders playing. We discuss how the cocreative shared value space for learning folk music was formed between the digital archive and the physical reality during our case study. We observed the roles the folk music players were taking and considering in this process of cultural transformation of folk music from the past. We point to the digital accessibility and gaps in passing forward regional music traditions through non-moderated archive mediation, and the need for a social and digital community loop around the original music artefacts in archives, to advance meaning-making and learning.

References

Bernats, Guntars, and Irena Trubina. 2017. 'Collective Music Making Challenges and Perspectives'. Journal of Pedagogy and Psychology Signum Temporis 9 (1): 1–6. https://doi.org/10.1515/sigtem-2017-0005.

Brattis, Pantelis, Emmanouel Garoufallou, Eugenios Politis, et al. 2024. 'A Greek Music Audiovisual Collections Platform: Presentation of the Open Source ReasonableGraph Platform for Music Collections'. Research Conference on Metadata and Semantics Research (Cham), 172–82.

Clough, Paul, Timothy Hill, Monica Lestari Paramita, and Paula Goodale. 2017. 'Europeana: What Users Search for and Why'. In Research and Advanced Technology for Digital Libraries, edited by Jaap Kamps, Giannis Tsakonas, Yannis Manolopoulos, Lazaros Iliadis, and Ioannis Karydis, vol. 10450. Lecture Notes in Computer Science. Springer International Publishing. https://doi.org/10.1007/978-3-319-67008-9_17.

Collins, Jez. 2018. 'Citizen Archiving and Virtual Sites of Musical Memory in Online Communities'. In The Companion to Popular Music History and Heritage. Routledge.

Davis, Jenny L. 2020. How Artifacts Afford: The Power and Politics of Everyday Things. Design Thinking, Design Theory. The MIT Press.

Grusin, Richard. 2015. 'Radical Mediation'. Critical Inquiry 42 (1): 124–48. https://doi.org/10.1086/682998.

Harrison, Rodney. 2013. Heritage: Critical Approaches. Heritage Studies. Routledge. https://doi.org/10.4324/9780203108857.

Hogsden, Carl, and Emma K Poulter. 2012. 'The Real Other? Museum Objects in Digital Contact Networks'. Journal of Material Culture 17 (3): 265–86. https://doi.org/10.1177/1359183512453809.

Honko, Lauri. 1998. 'Folklooriprotsess. Mäetagused'. Hüperajakiri 6. http://www.folklore.ee/tagused/nr6/honko.htm.

Horst, Heather A., and Daniel Miller. 2012. Digital Anthropology. Berg.

Hurley, Zoe. 2019. 'Imagined Affordances of Instagram and the Fantastical Authenticity of Female Gulf-Arab Social Media Influencers'. Social Media + Society 5 (1): 2056305118819241. https://doi.org/10.1177/2056305118819241.

Järvekülg, Madis, and Indrek Ibrus. 2022. ‘Auto-Communicative Reconstruction of Meaningfulness in Musical Randomness: Reclaiming Musical Order on Facebook’. Media, Culture & Society 44 (3): 549–73.

Kalkun, Mari. 2013. Pärimuse Ja Loomingu Suhestamine Pärimusmuusikaõpetuses. Loomingulise Magistrieksami Kirjalik Osa. Eesti Muusika- Ja Teatriakadeemia Ja Tartu Ülikooli Viljandi Kultuuriakadeemia Pärimusmuusika Ühisõppekava. Helsingi–Tallinn. https://core.ac.uk/download/pdf/79106976.pdf.

Keane, Webb. 2018. ‘Perspectives on Affordances, or the Anthropologically Real: The 2018 Daryll Forde Lecture’. HAU: Journal of Ethnographic Theory 8 (1–2): 27–38. https://doi.org/10.1086/698357.

Kelmendi, Arsim. 2024. ‘Sound Identity as a Phenomenon. Research on the Cultural Significance of Music in Ethnic and Subcultural Communities’. Interdisciplinary Cultural and Humanities Review 3 (2): 16–23.

Kivijärvi, Sanna, and Taru-Anneli Koivisto. 2025. ‘Ethical Questions in Transforming Music Practices’. Approaches: An Interdisciplinary Journal of Music Therapy 17 (1): 2025. https://doi.org/10.56883/aijmt.2025.611.

Kõmmus, Helen. 2023. ‘How to Participate in Participatory Music Making at a Contemporary Folk Music Festival: Runosong Nests at the Viljandi Folk Music Festival and Pelimanni. Evenings at the Kaustinen Folk Music Festival’. Folklore: Electronic Journal of Folklore 91: 141–64.

Lee, Brent. 2000. ‘Issues Surrounding the Preservation of Digital Music Documents’. Archivaria, 193–204.

Loomis, Jay. 2025. ‘3D Printing Reproductions of Indigenous Instruments in Museum Collections: Ethical Replication and Acoustic Sovereignty, Music & Science, 8’. https://doi.org/10.1177/20592043251364053.

Majsova, Natalija, and Jasmina Šepetavc. 2023. ‘Popular Music as Living Heritage: Theoretical and Practical Challenges Explored through the Case of Slovenian Folk Pop’. Int. J. Herit. Stud. 29 (12): 1283–98.

Markham, Annette N., and Gabriel Pereira. 2019. ‘Analyzing Public Interventions through the Lens of Experimentalism: The Case of the Museum of Random Memory’. Digital Creativity 30 (4): 235–56. https://doi.org/10.1080/14626268.2019.1688838.

Mucha, Franziska. n.d. Co-Creative Events for Engagement with Digital Cultural Heritage Collections.

Nagy, Peter, and Gina Neff. 2015. ‘Imagined Affordance: Reconstructing a Keyword for Communication Theory’. Social Media + Society 1 (2): 2056305115603385. https://doi.org/10.1177/2056305115603385.

Ngoma, Kutala, and Zoliswa Fikelepi-Twani. 2024. ‘Decolonizing the Teaching and Learning of Indigenous Nguni Music Instruments in Higher Institutions of Learning in South Africa. E-Journal of Humanities’. Arts and Social Sciences 5 (5): 2. https://doi.org/10.38159/ehass.2024552.

Paananen, Siiri, and Jonna Häkkilä. 2025. ‘Participatory Approaches to Ethical Design with Indigenous Cultural Heritage in the Digital Age’. In Digital Indigenous Cultural Heritage. Springer Nature Switzerland.

Pata, Kai. 2024. ‘Cultural Resilience Practices in the Digital Heritage Ecosystem’. KUI ’24: Proceedings of the 21th International Conference on Culture and Computer Science: From Humanism to Digital Humanities Article No, 1–6.

Purohit, Mekhala Vinod, and H. Bhavana. 2025. ‘Design of Chatbot (MusicBot) for Music Recommendation System Using Deep Learning’. 2025 6th International Conference on Inventive Research in Computing Applications (ICIRCA), 1835–42.

Raheb, Katerina El., Lori Kougioumtzian, Marina Stergiou, et al. 2022. ‘Designing an Augmented Experience for a Music Archive: What Does the Audience Need beyond the Sense of Hearing?’ ACM Journal on Computing and Cultural Heritage 15 (4): 1–24.

Ramnarine, Tina Karina. 1996. ‘Folk Music Education: Initiatives in Finland. – Folk Music Journal, 7, 2, 136–154’. Published By: English Folk Dance + Song Society. https://www.jstor.org/stable/4522543.

Ranjgar, Babak, Abolghasem Sadeghi, Maryam Shakeri, Fatema Rahimi, and Soo-Mi Choi. 2024. ‘Cultural Heritage Information Retrieval: Past, Present, and Future Trends’. IEEE Access 12: 42992–3026. https://doi.org/10.1109/ACCESS.2024.3374769.

Rehfeldt, Ruth Anne, Ian Tyndall, and Jordan Belisle. 2021. ‘Music as a Cultural Inheritance System: A Contextual-Behavioral Model of Symbolism, Meaning, and the Value of Music’. Behavior and Social 30 (1): 749–73.

Robinson, Kyle, and Dan Brown. 2021. ‘Quantitative User Perceptions of Music Recommendation List Diversity’. Proceedings of the 22nd ISMIR Conference, Online, November 7-12, 2021, 562–68.

Santos, Aandeline Dos, and Giorgos Tsiris. 2021. ‘Playing Marbles, Playing Music’. Approaches 13 (1): 3–5. https://doi.org/10.56883/aijmt.2021.153.

Särg, Taive. 2004. Mis on Eesti Rahvamuusika? – Mõeldes Muusikast. Sissevaateid Muusikateadusesse. Edited by Toim J. Ross and K. Maimets. Varrak.

Särg, Taive, and Ants Johanson. 2011. ‘Pärimusmuusika Mõiste Ja Kontseptsiooni Kujunemine Eestis’. Mäetagused 49: 115–38.

Sarv, Mari, Ave Goršič, and Risto Järv. 2023. ‘Performing and Archive: Aims, Interests, Ideologies and Expectations’. Folklore 91: 7–24.

Sengupta, Roshni. 2024. ‘Negotiating Minority Identities in Europe through Cultural Preservation: Music as Heritage among the Dutch Hindustani Diaspora in the Netherlands’. J. Intercultural Stud. 45 (1): 111–26.

Sinaga, Syahrul Syah, Restu Ayu Mumpuni, Fajry Sub’haan Syah Sinaga, Siti Aesijah, and Antonius Edi Nugroho. 2025. ‘Innovating Music Education Through Digital Platforms: A Mixed-Methods Approach to Enhancing Cultural Awareness’. Korean Music Education Society 54 (2): 139–60.

Six, Joren. 2021. ‘Panako 2.0: Updates for an Acoustic Fingerprinting System’. In Demo Session of the 22nd Int, edited by Extended Abstracts for the Late-Breakin. Society for Music Information Retrieval Conf. https://archives.ismir.net/ismir2021/latebreaking/000039.pdf.

Six, Joren, and Marc Leman. 2014. ‘Panako-A Scalable Acoustic Fingerprinting System Handling Time-Scale and Pitch Modification’. In15th International Society for Music Information Retrieval Conference (ISMIR 2014), 259–64.

Smolicz, Jerzy. 1981. ‘Core Values and Cultural Identity’. Ethnic and Racial Studies 4 (1): 75–90. https://doi.org/10.1080/01419870.1981.9993325.

Sugimoto, Shiego, Senan Kiryakos, Chiranthi Wijesundara, Winda Monika, Tetsuya Mihara, and Mitsuharu Nagamori. 2018. ‘Metadata Models for Organizing Digital Archives on the Web’. DCMI’18 (Dublin), 95–105.

Sun, Yinan, and Daniel D Suthers. 2023. ‘From Affordances to Cultural Affordances: An Analytic Framework for Tracing the Dynamic Interaction among Technology, People and Culture’. Cultures of Science 6 (2): 235–49. https://doi.org/10.1177/20966083231176863.

Tang, Zhe, and Ennan Wu. 2025. ‘Polyphony in Action: Interdisciplinary Approaches and Sustainable Strategies for Musical Cultural Heritage Preservation’. Humanities and Social Sciences Communications 12 (1): 1–12.

Taylor, Joel, and Laura Kate Gibson. 2017. ‘Digitisation, Digital Interaction and Social Media: Embedded Barriers to Democratic Heritage’. International Journal of Heritage Studies 23 (5): 408–20. https://doi.org/10.1080/13527258.2016.1171245.

Terras, Melissa, Stephen Coleman, Steven Drost, et al. 2021. ‘The Value of Mass-Digitised Cultural Heritage Content in Creative Contexts’. Big Data & Society 8 (1): 20539517211006165. https://doi.org/10.1177/20539517211006165.

Urbancov, Hana. 2024. ‘Sound Recordings of Folk Music at the ü of Musicology, Slovak Academy of Sciences’. Musicologica Slovaca 15 (41): 252–75. https://doi.org/10.31577/musicoslov.2024.2.6.

Villaespesa, Elena. 2019. ‘Museum Collections and Online Users: Development of a Segmentation Model for the Metropolitan Museum of Art’. Visitor Studies 22 (2): 233–52. https://doi.org/10.1080/10645578.2019.1668679.

Waysdorf, Abby S. 2021. ‘Remix in the Age of Ubiquitous Remix’. Convergence: The International Journal of Research into New Media Technologies 27 (4): 1129–44. https://doi.org/10.1177/1354856521994454.

14:30–15:30 *PANEL* **[51]**

Uncovering Hidden Infrastructures for Digital Humanities in GLAMs

Mahendra Mahey[1], Pille Pruulmann-Vengerfeldt[2], Hans Dam Christensen[3], Berndt Clavier[4], Rikke Lie Halberg[5], Paula Bray[6]

[1] *Tallinn University,University of Strathclyde,Estonian National Museum*
[2] *Malmö University,Estonian National Museum*
[3] *University of Copenhagen*
[4] *Malmö University*
[5] *University of Lund*
[6] *State Library of Victoria*

Keywords: *Digital Humanities Research, Laboratories, Incubation, Digital Transformation, Infrastructures*

How GLAMs are infrastructures for digital humanities - what kind of discussions are emerging and missing from these collaborations?

Digital Humanities (DH) research over the last few decades has increasingly depended on Galleries, Libraries, Archives, and Museums (GLAMs) especially in the context of accessing the digital cultural heritage these organisations curate and manage. GLAM institutions are therefore often framed as merely service providers or data sources and rather overlooked as research infrastructures in their own right, ones that actively shape what kinds of knowledge, methods, and collaborations become possible in DH scholarship. This panel argues that GLAMs function as socio-technical research infrastructures: they organise access to collections, define standards and metadata practices, provide labs and experimentation spaces, and often structure their work and organisational culture to form partnerships

and collaborations establishing some of the foundations of much DH work. Making these GLAM infrastructures visible, recognised, acknowledged and valued is essential for understanding both the possibilities, inequalities and gaps in supporting digital scholarship in different contexts.

The panel brings together three international comparative examples from Northern Europe and Australia to demonstrate how GLAM infrastructures enable, restrict, and reorient DH practice across different organisational contexts and at different scales. Rather than presenting isolated initiatives, the session analyses aspects of research infrastructures such as GLAM Labs, incubators, repositories and catalogues as elements that contribute to the production of specific research cultures and mindsets. By comparing small and medium Estonian museums, Nordic collaborative infrastructures in the Öresund region, and the State Library of Victoria Lab in Australia, the panel highlights shared challenges around sustainability, recognition, interoperability, and equity.

Methodological approach

The panel adopts a comparative, practice-based methodology combining:

institutional examples,

reflective practitioner accounts,

analysis of organisational processes and infrastructures,

and cross-case synthesis during the panel discussion.

This approach highlights infrastructures as living and evolving systems rather than purely technical platforms, drawing on recent work in cultural infrastructure studies and digital heritage research.

Across the examples, we ask:

How do GLAM infrastructures shape DH research practices and agendas?

What forms of collaboration, mindsets, cultures, care, and maintenance are required but often invisible in DH research taking place with GLAMs?

Which infrastructural gaps persist (skills, funding, sustainability, recognition)?

What models might support more equitable and durable partnerships between GLAMs and DH researchers?

The panel sheds light on what there is and what is missing and advances a more reflexive understanding of GLAMs not simply as repositories of heritage but as co-producers of knowledge and research. Recognising GLAMs as infrastructures enables Digital Humanities to move beyond project-based experimentation toward sustainable, socially responsible, and institutionally embedded forms of digital scholarship.

Case contributions

Museum Innovation Infrastructures in Estonia: Incubators/Labs as DH Capacity Builders

We examine how Estonian museum incubators/labs can be infrastructure for innovation and collaboration in supporting and developing Digital Humanities research and practice. Incubators are not merely funding mechanisms or technology accelerators, but critical infrastructures for cultural experimentation, capacity-building, and social transformation across the GLAM and DH sectors. The discussions are partly informed by the work of the DOORs project (Digital Incubator for Museums) that ran for two years from 2021-2023.

Through critical reflections and discussions with practitioners and leaders in Estonian GLAMs and cultural heritage, this panel describes the extent to which small and medium-sized museums engage with digital maturity models, community-driven innovation, sustainability incubator frameworks and DH research. It highlights how these infrastructures can enable museums and digital humanities researchers to experiment with AI, data-driven storytelling, and participatory design while navigating resource constraints, policy frameworks, and ethical challenges.

The panel will critically ask how incubators can reconfigure museums as digital humanities partners and what kinds of knowledge, partnerships, and strategies are being produced or could be developed to

work with DH research. What conversations about equity and long-term support remain missing? The discussion repositions incubators as both catalysts and case studies for rethinking the infrastructures for digital culture and digital humanities research.

Mahendra Mahey, Estonian National Museum (Junior Researcher), Tallinn University (Research Data Analyst (Humanities)) and University of Strathclyde (PhD Candidate)

Digital heritage for democracy theme day discussion from the Centre of Modern European Studies (CEMES) collaboration

As GLAMs raise resources to digitalise and make heritage accessible online, they are a crucial part of the infrastructure for the digital humanities. There is also a broader question: how are digital resources an infrastructure for democracy? The presentation looks back at a series of discussions between museum, library and archive professionals and students from Malmö, Lund and Copenhagen universities who study issues related to heritage and cultural institutions. The professionals took with them a digitalisation-related problem or a challenge, and student groups moved around the discussion.

Questions and problems that arose from the professionals related to AI as a digitalisation tool, longevity (or quick obsoleteness) of the technologies, communication and awareness-raising, interoperability between different Nordic countries, and questions of how to address the lack of representativeness in the collections. All of these problems have strong implications to Digital Humanities research. With today's level of trustworthiness and reliability of AI tools, how will using them to register collections affect the quality and value of the research that can later be produced on these collections? How is the long-term sustainability of the digital infrastructures affected by the fast-paced technological change?

The discussions and their outcomes will be analysed using TRUST principles for digital repositories by Lin et al (2020) where the acronym stands for Transparency, Responsibility, User focus, Sustainability and Technology. The paper looks at how these principles become challenges in practice for GLAMs when they work with digital repositories.

Pille Pruulmann-Vengerfeldt, Professor of Media and Communication, Malmö University, visiting researcher at the Estonian National Museum.

Hans Dam Christensen, Professor in Cultural Communication, Department of Communication, University of Copenhagen.

Berndt Clavier, Senior Lecturer in the School of Arts and Communication at Malmö University.

Rikke Lie Halberg, PhD student in History, University of Lund.

What Is There and What Is Missing: GLAM Labs as Infrastructures for Digital Humanities – Reflections from the State Library of Victoria (SLV) Lab?

The State Library Victoria Lab as an example of a GLAM-embedded digital experimentation space. Guided by principles of process, publishing, and people, the Lab activates collections through open datasets, creative reuse, and collaborative research. Including this Australian case provides a useful counterpoint to the Northern European contexts: it offers a larger-scale model that highlights issues of sustainability, recognition of invisible work, and institutional embedding that some smaller European initiatives are now confronting. The comparison demonstrates that infrastructural challenges transcend geography and scale, strengthening the panel's broader position about GLAMs as global DH infrastructures.

Bringing together curators, technologists, and researchers working with SLV Lab and comparable initiatives, this panel examines both what is there, the tools, datasets, spaces, and communities enabling digital experimentation, and what is missing: recognition of the work needed, critical frameworks for impact assessment, and enduring strategic models of collaboration between GLAMs and the Digital Humanities.

Through specific contributions, the discussion situates GLAM Labs as laboratories of cultural knowledge-making and explores how their hybrid infrastructures can evolve from serving Digital Humanities projects to actively shaping them, rethinking not just access to cultural data, but the values and practices that sustain it.

Paul Bray, Chief Digital Officer, State Library of Victoria, Melbourne, Australia.

Session Format (60 minutes)

3 × 10 min short provocations

15 min moderated cross-case dialogue

15 min audience contributions, discussion and synthesis

The session prioritises exchange between practitioners and researchers, inviting participants to reflect on their own institutional infrastructures and identify transferable strategies.

Contribution to the field

By reframing GLAMs as infrastructures rather than service providers, this panel contributes:

a comparative vocabulary for analysing DH–GLAM relations,

case studies / examples,

practical insights for building sustainable, equitable digital research ecosystems.

to move discussions of Digital Humanities from projects to infrastructures, from tools to maintenance, and from access to responsibility.

References

"Balanced Value Impact Model: The Balanced Value Impact Model." Balanced Value Impact Model, https://www.bvimodel.org/bvim/index. Accessed 29 October 2025.

Christensen, Clayton M. The Innovator's Dilemma: When New Technologies Cause Great Firms to Fail. Paperback, Harvard Business Review Press, 2016.

DOORS - Digital Incubator for Museums. https://ars.electronica.art/doors/en/. Accessed 29 October 2025.

Giglitto, Danilo, et al., editors. Cultural Heritage and Social Impact: Digital Technologies for Social Inclusion and Participation - Symposium Companion. Sheffield Hallam University, 2021.

"Library Labs as Experimental Incubators for Digital Humanities Research." TPDL 2019, 23rd International Conference on Theory and Practice of Digital Libraries, Abstracts, 2019, http://hdl.handle.net/1854/LU-8645483.

Lin,D. et al. 2020. 'The TRUST Principles for Digital Repositories'. Scientific Data 7(1):144. doi:10.1038/s41597-020-0486-7.

Mahey, M., Al-Abdulla, A., Ames, S., Bray, P., Candela, G., Chambers, S., Derven, C., Dobreva-McPherson, M., Gasser, K., Karner, S., Kokegei, K., Laursen, D., Potter, A., Straube, A., Wagner, S-C. and Wilms, L. with forewords by: Al-Emadi, T. A., Broady-Preston, J., Landry, P. and Papaioannou, G. (2019) Open a GLAM Lab. Digital Cultural Heritage Innovation Labs, Book Sprint, Doha, Qatar, 23-27 September 2019.

Nowviskie, Bethany. "A Skunk in the Library." Bethany Nowviskie, 28 June 2011, https://nowviskie.org/2011/a-skunk-in-the-library/. Accessed 29 October 2025.

Pereda, Javier, et al. "Online Cultural Heritage as a Social Machine: A Socio-Technical Approach to Digital Infrastructure and Ecosystems." International Journal of Digital Humanities, Mar. 2025, https://doi.org/10.1007/s42803-025-00097-6.

Tanner, Simon. Delivering Impact with Digital Resources: Planning Strategy in the Attention Economy. Facet Publishing, 2020.

Terras, Melissa, James Baker, et al. "Enabling Complex Analysis of Large-Scale Digital Collections: Humanities Research, High-Performance Computing, and Transforming Access to British Library Digital Collections." Digital Scholarship in the Humanities, vol. 33, no. 2, June 2018, pp. 456–66, https://doi.org/10.1093/llc/fqx020.

Terras, Melissa. "Opening Access to Collections: The Making and Using of Open Digitised Cultural Content." Online Information Review, edited by Professor G.E. Gorman And Professor Jennifer Rowley, vol. 39, no. 5, Sept. 2015, pp. 733–52, https://doi.org/10.1108/OIR-06-2015-0193.

Terras, Melissa, Stephen Coleman, et al. "The Value of Mass-Digitised Cultural Heritage Content in Creative Contexts." Big Data & Society, vol. 8, no. 1, 2021, p. 20539517211006164, https://doi.org/10.1177/20539517211006165.

Terras, Melissa M. "The Rise of Digitization." Digitisation Perspectives, edited by Ruth Rikowski, SensePublishers, 2011, pp. 3–20, https://doi.org/10.1007/978-94-6091-299-3_1.

The Social Impact of Cultural Heritage Was Discussed at the Seminar Dedicated to the Faro Convention | Kultuuriministeerium. https://www.kul.ee/en/news/social-impact-cultural-heritage-was-discussed-seminar-dedicated-faro-convention. Accessed 29 October 2025.

The State Library of Victoria Lab. https://lab.slv.vic.gov.au/. Accessed 29 October 2025.

Thylstrup, Nanna. The Politics of Mass Digitization: PhD Thesis. 2014, https://research.cbs.dk/en/publications/the-politics-of-mass-digitization-phd-thesis.

User Studies For Digital Library Development. https://alastore.ala.org/content/user-studies-digital-library-development. Accessed 29 October 2025.

Vershbow, Ben. “NYPL Labs: Hacking the Library.” Journal of Library Administration, vol. 53, no. 1, Jan. 2013, pp. 79–96, https://doi.org/10.1080/01930826.2013.756701.

What Is Digital Humanities and What’s It Doing in the Library? – In the Library with the Lead Pipe. 27 June 2012, https://www.inthelibrarywiththeleadpipe.org/2012/dhandthelib/.

Wilms, Lotte, et al. Europe’s Digital Humanities Landscape: A Study From LIBER’s Digital Humanities & Digital Cultural Heritage Working Group. Zenodo, 17 June 2019, https://doi.org/10.5281/zenodo.3247286.

Wittmann, Rachel, et al. “From Digital Library to Open Datasets: Embracing a ‘Collections as Data’ Framework.” Information Technology and Libraries, vol. 38, no. 4, Dec. 2019, pp. 49–61, https://doi.org/10.6017/ital.v38i4.11101.

Session 3A — 13:30–15:10

13:30–13:50 **From Old News to New Tools**
Johan Heinsen, Matias Kokholm Appel

13:50–14:10 **Tracing Industrial Modernity in Global Historical Newspaper Collections using LLMs**
Sophie-Marie Ertelt, Muhammad Okky Ibrohim, Andres Karjus

14:10–14:30 **Reanimating early Danish periodical journals (1740-1770) as digital text: Practices of historical transcription and loss of materiality**
Maria Nørby Pedersen

14:30–14:50 **Exploring AI-Supported Qualitative Data Analysis**
Daniel Andersson

14:50–15:10 **Using Large Language Models for searching explainable relations in a cloud of Cultural Heritage knowledge graphs: SampoSampo as a neuro-symbolic system**
Annastiina Ahola, Petri Leskinen, Heikki Rantala, Jouni Tuominen, Eero Hyvönen

13:30–13:50 *SHORT PAPER* **[52]**

From Old News to New Tools

Johan Heinsen, Matias Kokholm Appel
Aalborg University, Denmark

Keywords: *History, Language modelling, BERT, Digital Humanities, Newspapers*

We introduce a language model designed to help researchers navigate the contextual and temporal domains of Danish and Norwegian newspaper sources until the end of absolutism in Denmark (1849). Newspapers dealt in a broad range of topics, but often presented these in a somewhat limited number of highly codified formats, many of which were specific to the period. The model is tested both qualitatively and quantitatively, with a special focus on identifying the boundaries of its domains.

13:50–14:10 *SHORT PAPER* **[53]**

Tracing Industrial Modernity in Global Historical Newspaper Collections using LLMs

Sophie-Marie Ertelt[1,2], Muhammad Okky Ibrohim[1], Andres Karjus[1,3,4]
[1] *University of Tartu, Estonia*
[2] *Örebro University, School of Business*
[3] *Estonian Business School*
[4] *Tallinn University, School of Humanities*

Keywords: *Industrial modernity, Large language models (LLMs), Text mining, Historical newspaper analysis, Environmental discourse*

The past two centuries have witnessed a profound transformation in how industrial societies imagine and enact the relationship between humans, nature, and technology (Beck, 1992; Latour, 2012). This transformation can be captured by the notion of industrial modernity, defined as a historically accumulated set of ideas, institutions, and practices that has structured social, economic, and environmental developments since the late eighteenth century. It has been primarily driven by two beliefs: that the natural environment can be separated from society and subordinated to human purposes, and that continuous scientific and technological progress will secure material improvement and overcome natural limits (Kanger et al., 2022, 2023). These assumptions have guided two centuries of industrial development, while simultaneously generating the ecological and social challenges of the present, including, but not limited to, climate change, biodiversity loss, and widening social inequality. Understanding how these underlying beliefs emerged, stabilised, and may have changed over time is thus essential for explaining why current unsustainable industrialisation trajectories persist despite mounting evidence of their limits (Kanger et al., 2022; Kanger & Schot, 2019).

Recent research has begun to operationalise the concept of industrial modernity by developing a multi-method approach for measuring its long-term evolution across multiple countries and technological, institutional, and ideational dimensions (Kanger et al., 2022, 2023; Ertelt et al., 2025). Text mining digitised newspaper collections has proven a particularly valuable source for capturing the public articulation of ideas about the environment, science, and technology (EST), the triad of core aspects of industrial modernity. As one of the most continuous and socially embedded records of communication (though, of course, not a perfect societal reflection), newspapers reveal how industrial societies have narrated technological progress and its relationship to nature. Yet the abundance and heterogeneity of historical texts also pose analytical challenges. Commonly used keyword frequency statistics only capture how often something is discussed but not how it is framed; fixed lists cannot follow linguistic and semantic change across time; and basic sentiment analysis, although popular, may compress complex rhetorical positions into binary positive–negative polarity.

Recent developments in generative large language models (LLMs) have opened up new avenues for analyzing complex beliefs and stances at scale in the humanities and social sciences, without requiring separate training for each research question or analytical variable (Karjus, 2025; Ziems et al., 2024). Here, we use an LLM-driven text-mining approach to trace the long-term evolution of the ideational dimension of industrial modernity in the Anglosphere — United Kingdom, United States, Canada, Australia, and New Zealand — between 1900 and 2025. LLMs are particularly suited to this task because they can parse meaning and stance in context, adapt to historical variation in vocabulary, and be guided by explicit annotation protocols, bridging the gap between close reading and large-scale analysis, where the final interpretation of the quantitized results remains the role of the human researcher. Here, we first detect whether each text extract matching an initial keyword search is relevant as an expression of EST, and classify whether it expresses one of the ideas of the separation of nature and society, the instrumental conception of nature as a stock of resources for human use, technology as the ultimate solution, or limitlessness or easy substitutability of natural inputs. We also codify frames or stances of how these ideas are discussed: e.g. neutrally, with criticism, or doubt, such as limits-to-growth arguments or scepticism toward technological optimism. Aggregated over time, these labels can be operationalized as country-specific and weighted global time series that trace the coevolution of ideas and real-world events, allowing us to map the continuity and potential ruptures of the ideational dimension of industrial modernity across the Anglosphere. We also evaluate LLM performance and discuss accounting for error rates in the analysis. Here we use the GPT-5.x family of models via the Proquest TDM platform (which provides full access to data but sets constraints on model use).

This work serves as a pilot study for a larger ongoing project of mapping a larger array of ideas across a larger set of G20 countries (supported by multilingual models such as the aforementioned), and contributes to the conference theme and broader DH debates around navingating overwhelmingly large textual corpora without losing interpretive nuance. While previous research has often portrayed long-term ideational change through counts of environmental or technological keywords, the LLM-based framework measures the worldviews and evaluative orientations embedded in relevant textual context. It shifts the analytical focus from the volume of discourse to ideas and socio-political stances, offering a new empirical basis for understanding how ideas of progress, control, and substitution have been linguistically sustained or contested over 125 years of public communication. At the same time, the study invites reflection on the epistemic condition of abundance itself: both the digital abundance

confronting today's scholars and the historical ideology of abundance that lies at the heart of industrial modernity, the conviction that natural resources and technological progress are inexhaustible.

References

References

Beck, U. (1992). Risk society: Towards a new modernity (Vol. 17). sage.

Ertelt, S.-M., Šeļa, A., Romanov, B., Ibrohim, M. O., Kanger, L., Maurer, L., Tinits, P., Velmet, A., & Lyly, N. (2025). Sustainability front-runners? Comparing the Evolution of Industrial Modernity in the Nordic-Baltic Region. Paper presented at the International Sustainability Transitions Conference 2025, Lisbon.

Kanger, L., Tinits, P., Pahker, A.-K., Orru, K., Tiwari, A. K., Sillak, S., Šeļa, A., & Vaik, K. (2022). Deep Transitions: Towards a comprehensive framework for mapping major continuities and ruptures in industrial modernity. Global Environmental Change, 72, 102447. https://doi.org/10.1016/j.gloenvcha.2021.102447

Kanger, L., Tinits, P., Pahker, A.-K., Orru, K., Velmet, A., Sillak, S., Šeļa, A., Mertelsmann, O., Tammiksaar, E., Vaik, K., Penna, C. C. R., Tiwari, A. K., & Lauk, K. (2023). Long-term country-level evidence of major but uneven ruptures in the landscape of industrial modernity. Environmental Innovation and Societal Transitions, 48, 100765. https://doi.org/10.1016/j.eist.2023.100765

Karjus, A. (2025). Machine-assisted quantitizing designs: Augmenting humanities and social sciences with artificial intelligence. Humanities and Social Sciences Communications, 12(1), 277. https://doi.org/10.1057/s41599-025-04503-w

Latour, B. (2012). We have never been modern. Harvard university press.

Ziems, C., Held, W., Shaikh, O., Chen, J., Zhang, Z., & Yang, D. (2024). Can Large Language Models Transform Computational Social Science? Computational Linguistics, 50(1), 237–291. https://doi.org/10.1162/coli_a_00502

14:10–14:30 *SHORT PAPER* **[54]**

Reanimating early Danish periodical journals (1740-1770) as digital text: Practices of historical transcription and loss of materiality

Maria Nørby Pedersen

Aarhus University, Denmark

Keywords: *digital history, enlightenment, periodicals, digital corpus, danish history*

Digital analysis requires digital text, and while transforming documents into digital texts is blooming, most of Danish history is currently inaccessible to the digital humanities (DM) due to the lack of sufficient open access data.[1] This paper presents a digitized corpus of 18th-century journals, a neglected part of Enlightenment literature that has remained outside the literary canon. The material has never been transformed into digital text before. It is the aim here to describe how digitalisation has reanimated the journals and to discuss the consequences of methodological choices which shape the possibilities of future analysis on the corpus. Digital historians agree that digitalisation transforms the historical documents (Jarlbrink and Snickars 2017) and creates new blind spots in our knowledge of the sources (Putnam 2016). Providing a description of the corpus construction is an important step to ensure a critical awareness of the provenance and transformation of the new digital sources. This paper therefore both contributes to DM with a new corpus and a discussion on the loss of materiality during transcription which shapes the digital sources by which we can access history in new ways.

Periodicals were the essence of 18th-century European Enlightenment (Pettegree 2014). While periodicals are central to English Enlightenment studies, the Danish canon, however, privileges Ludvig Holberg as its sole representative of the period (Schön 2020). He is considered the father of Danish fiction and his legacy is only slightly balanced by a few early modern and enlightenment writers on the optional list supplementing the canon (Kanonudvalgets rapport 2025). However, research has shown that periodicals played a crucial role in shaping the Nordic Enlightenment (Krefting et al. 2015). The early period of periodicals (1740-1770) was particularly distinctive and therefore relevant to widening our view of the literary landscape (Krefting 2018). This project seeks to make these journals, which have been lost in the abundance of 18th-century periodical writings and forgotten in later debates about the Danish Enlightenment, digitally accessible and analytically tractable.

The corpus consists of 24 Danish-language journals (1740-1770). The selection of material has been made according to the loosely defined boundaries between newspapers, small prints and periodical

journals made by researchers (Tjønneland 2008; Nøding 2017) as well as research interests of this project to study the textual layer.[2] The selected journals comprise a total of 40 physical volumes, which have been digitalized as pdf by the Danish and Norwegian Royal Libraries in previous projects and as part of the Danish Royal Library's "digitalization on demand" solution. The corpus has been transcribed with AI text recognition (TR) in *Transkribus*. The work was done by the author of this paper in four phases. Phase 1 consisted of exploring the functionalities and cost-benefit of field, layout and TR models in January 2025. Field recognition was abandoned because of the cost, and while there existed a few TR models built on Danish newspapers, these were not suitable for the journals. The TR faults were probably due to differences between the newspaper models' training data and the journals. Newspapers were typically typeset with a small gothic script in columns; they were digitalized mostly as black-and-white copies, and they were of a slightly later period. The journals resemble books, with one main text and larger font, and they were digitalised as colour images. At the time it was therefore more effective to adopt a flexible workflow with models trained on the corpus itself. Phase 2 consisted of training layout models; Phase 3 of training TR models and transcribing (February – April 2025). The final TR model arrived at 0.29% accuracy. It was trained with *Avisfraktur 0.00006a* by Johan Heinsen as the base model on 972 pages (245.265 words) and with binarization on April 13, 2025. Phase 4 involved manual correction of individual pages with a focus on front pages where recognition often failed on titles and first letters, as well as pages with complex layout. The corpus will be exported as txt files and published on loar.kb.dk in 2026, ensuring openness, continued preservation and future use of the corpus beyond the original project's scope (see Graham et al. 2022, xvi, 38, 49).

All digitalization choices have been guided by the need to have a tractable digital text in the end with as few messy elements as possible. The layout analysis was deliberately shaped to disregard headings, footings, folio numbers, pictures, vignettes and catchwords. The construction has been led by the aim to analyse textual qualities, not aspects of e.g. book history making the loss of materiality (e.g. size, binding, printing and pictures) important to consider. Likewise, the corpus is not apt for linguistic analysis since the TR of punctuation is poor. One consequence of the poor punctuation TR is the loss of evidence of the practice of inserting characters such as *** instead of names hindering character network analysis. The messiness of number recognition has also been disregarded, since numbers are often cleaned during pre-processing. The downside of this is that tables are not transcribed sufficiently in the corpus, and the page numbering is often wrong. Other recurring TR mistakes to be aware of are confusion between i/j, f/s, h/b and u/n.

DM is not just about the tools, but also about understanding what "the digital does and has done to our understanding of the past and ourselves" (Graham et al. 2022, 23). Researchers have highlighted that uncritical digitalisation and datafication of historical sources naturalise human thought on a narrow sample of the past's voices (i.e. Hitchcock 2012). One way to counter this change is by remembering that data are still sources we need to critically question (by whom and how are the (digital) texts created), and review physically if possible, before drawing conclusions about the past (Putnam 2016; Heinsen 2023).

[1] Existing projects: mime-memo.github.io; hislab.quarto.pub/eno/; kb.dk/inspiration/trykkefrihedens-skrifter.

[2] The corpus was made as part of a postdoc project (2025-26) in the Velux-funded project "The Rise of Science and Fiction during the Danish Enlightenment", PI Simona Zetterberg-Nielsen, Aarhus University.

References

Graham, Shawn, Ian Milligan, Scott B. Weingart, and Kim Martin. 2022. Exploring Big Historical Data. The Historians Macroscope. 2nd ed. World Scientific.

Heinsen, Johan. 2023. 'Kilde Og Data: Overvejelser Om Historiefaget Og de Digitale Metoder'. Temp - Tidsskrift for Historie 13 (26): 186–96.

Hitchcock, Tim. 2012. 'A Five Minute Rant for the Consortium of European Research Libraries'. Historyonics, October 29. historyonics.blogspot.com/2012/10/a-five-minute-rant-for-consortium-of.html.

Jarlbrink, Johan, and Pelle Snickars. 2017. 'Cultural Heritage as Digital Noise: Nineteenth Century News-papers in the Digital Archive'. Journal of Documentation (BINGLEY) 73 (6): 1228–43.

Kanonudvalgets rapport 2025.

Krefting, Ellen. 2018. ‘News versus Opinion: The State, the Press, and the Northern Enlightenment’. In Travelling Chronicles: News and Newspapers from the Early Modern Period to the Eighteenth Century, edited by Siv Gøril Brandtzæg, Paul Goring, and Christine Watson. Brill.

Krefting, Ellen, Aina Nøding, and Mona Ringvej, eds. 2015. Eighteenth-Century Periodicals as Agents of Change. Perspectives on Northern Enlightenment. Brill.

Nøding, Aina. 2017. Periodical Fiction in Denmark and Norway before 1900. https://doi.org/10.1093/acrefore/9780190201098.013.293.

Pettegree, Andrew. 2014. The Invention of News. How the World Came to Know about Itself. Yale Univer-sity Press.

Putnam, Lara. 2016. ‘The Transnational and the Text-Searchable: Digitized Sources and the Shadows They Cast’. The American Historical Review (OXFORD) 121 (2): 377–402. https://doi.org/10.1093/ahr/121.2.377.

Schön, Theresa. 2020. A Cosmography of Man: Character Sketches in ‘The Tatler’ and ‘The Spectator’. De Gruyter.

Tjønneland, Eivind, ed. 2008. Opplysningens Tidsskrifter. Norske Og Danske Periodiske Publikasjoner På 1700-Tallet. Fagbokforlaget.

14:30–14:50 *LONG PAPER* **[55]**

Exploring AI-Supported Qualitative Data Analysis

Daniel Andersson

Umeå University, Sweden

Keywords: *Qualitative Data Analysis, AI, Methodology*

Exploring AI-Supported Qualitative Data Analysis

The rapid development of generative AI has brought to the forefront important questions concerning computer-assisted qualitative data analysis (CAQDAS). While the advantages of digital tools for quantitative analysis are widely recognized, their implications for deeply qualitative research remain less clear.

This exploratory study examines the use of generative AI in qualitative analysis through both AI-integrated CAQDAS software, specifically *ATLAS.ti*, and a conversational AI tool, *ChatGPT Pro*. A small corpus of previously undigitized texts is analyzed with and without AI assistance. As the researcher is personally familiar with the contextual background of the material, the study allows for a grounded comparison between traditional qualitative interpretation and AI-supported analysis.

The text corpus consists of a short weekly newsletter published internally within a university department at Umeå University between 1977 and 1982. The newsletter contains information relevant to the field of Scandinavian languages and it has the potential to shed light on the development of the academic discipline in question, as well as pedagogical and organizational changes within a Swedish university.

The project discusses potential advantages and pitfalls of AI involvement in qualitative research, addressing issues of ethics, reliability, and added value. It also reflects on how generative AI potentially challenges and reshapes the very nature of qualitative analysis as a method.

14:50–15:10 *SHORT PAPER* **[56]**

Using Large Language Models for searching explainable relations in a cloud of Cultural Heritage knowledge graphs: SampoSampo as a neuro-symbolic system

Annastiina Ahola[1], Petri Leskinen[1], Heikki Rantala[1], Jouni Tuominen[2,3,1], Eero Hyvönen[1,3]

[1] *Aalto University, Department of Computer Science, Finland*

[2] *University of Helsinki, Helsinki Institute for Humanities and Social Sciences (HSSH), Finland*

[3] *University of Helsinki, Helsinki Centre for Digital Humanities (HELDIG), Finland*

Keywords: *linked data, knowledge graphs, cultural heritage, relational search, relation extraction*

Knowledge discovery of “interesting” or even serendipitous relations in data, often called relational search, provides a novel Artificial Intelligence-based approach in Digital Humanities for studying Cultural Heritage. Relational search methods are traditionally symbolic, based on searching connections in knowledge graphs. In contrast, this paper presents a novel neuro-symbolic approach to relational search based on combining Large Language Models (LLM) with knowledge graphs (KG). It is argued that by using curated KG data and data models with Retrieval-Augmented Generation, hallucinations of LLMs can be mitigated and relational search extended also to web resources external to the underlying cloud of KGs. As a practical use case, first results of using the method for knowledge discovery as part of the new web service *SampoSampo – Connecting Everything to Everything Else* are presented.

Session 3B — 13:30–15:20

13:30–13:50 **How About a Game of Sáhkku? How digitizing an Archive Empowered Treasure Hunting**
Therese Foldvik, Ida Tolgensbakk

13:50–14:10 **Appraising Archival Icebergs: Digitizing the Archive of the Fogelstad College for Women’s Political and Civic Rights and Duties**
Christa Shusko

14:10–14:30 **Fragments, bias, lost voices and measuring silence in the digitised archive: the case of early modern maritime disability**
Catherine Beck

14:30–14:50 **The urge to digitise: Medieval materials in museum and archival collections**
Olga Zabalueva, Polina Ignatova

14:50–15:20 **A Best-Practice Pipeline for Nuanced, Reproducible Bibliographic Data Science -- Case VD17**
Eetu Mäkelä, Thea Lindquist

13:30–13:50 *SHORT PAPER* **[57]**

How About a Game of Sáhkku? How digitizing an Archive Empowered Treasure Hunting

Therese Foldvik[1], Ida Tolgensbakk[2]
[1] *IKOS, University of Oslo, Norway*
[2] *Stiftelsen Norsk Folkemuseum, Norway*

Keywords: *archives, infrastructure, transcription, board games*

This paper will present how a large research infrastructure for tradition archives led to unexpected discoveries in our collections: realising that we had items and collections in our archive that nobody had seen (or cared to look for) for many decades. The technical procedure of digitizing archival material turned out to also be a valuable basis for knowledge production. We will use an example from the Norwegian Folk Museum, to answer the question: What types of knowledge production and what kind of knowledge does a digitisation process generate?

With the digitization of the three largest tradition archives in Norway through the SAMLA research infrastructure project, important collections of tales and legends, songs and folk belief are now made accessible and inter-searchable. Norsk Folkemuseum contributed with the full series of questionnaires sent from the Norwegian Ethnological Research from 1947 to 2016, on topics ranging from traditional agriculture through brewing and celebrations to histories about the modern tooth fairy.

When searching SAMLA, users browse through the original records that folklore collectors in the 19th and 20th centuries wrote down during their travels. It is also possible to explore systematic and schematic collections, to map customs, beliefs, and work from around the country. We, the scholars, archivists and technical staff in SAMLA knew the archives well. We knew that not all parts of the collections were well-organised, and that a large part of the digitising process would have to entail ordering the physical archive to make the items ready for digitization – and not surprisingly, while

working on ordering, scanning, and annotating these materials, lesser-known materials were also unearthed. One such discovery was the document hidden in archive number NEG Varia 11349: an original manuscript from the amateur folklorist Yngvar Mejland. He collected information on the Sami board game Sáhkku in the 1940s. Sáhkku is a dice game played for two players, known from large parts of the Sami area, from Lule Sami in the south to Skolt Sami in the northeast, and has existed in many local variants (Berg-Nordlie 2019). The game was almost – but not completely – out of living tradition when enthusiasts in recent times took steps to revitalize it (Berg-Nordlie & Tolgensbakk 2024). The joy of rediscovering Mejlands original has prompted the archive and the museum it belongs to into finding ways of disseminating knowledge of the board game, and will be doing so in collaboration with Sami organisations and individuals in the coming years.

In our paper, we will use the example of the Sáhkku board game to ask: What kind of knowledge does a digitisation process generate? We want to point to how the discoveries has created new opportunities in our research fields and for the archives themselves. At least two types of knowledge are in play, and we will touch on both of them:

- What knowledge is needed to digitise: what is this and how can we digitise it?
- What knowledge is created about the digitised object and the system it is part of, when the objects are gathered, handled, and systematised in the new infrastructure?

The first type of knowledge is what we looked for; in a digitization process, the goal is often to uncover information about the material that makes it easier to work with and ensures a good digital product (Foldvik 2026). This can e.g. be the condition of the document and how to handle it. Can it be unfolded to photograph the text? Does it contain any basic information as author, date and location which can be used as metadata? If making a digital transcription; what language and type of handwriting is it in?

The second type of knowledge is what we didn't look for but emerges in the process. Working closely with a text will naturally often uncover new or unexpected information. When investigating a document, our eyes fall upon small details which makes us compare it to other material or knowledge we have about a phenomenon, location or person, or is completely new insights. Even though this can be referred to as a secondary knowledge production, the information is far from secondary, but can uncover a whole new insight to archives, documents and context (Foldvik 2026). When digitizing an archive, it raises the awareness of people working with the material. One of the things that might happen is that we now know more of what our archive contains. In our day-to-day work, previous knowledge we have gained from working with the material may sometimes trigger curiosity. And before you realize it you dive into the digital material, flipping through images or searching through the metadata. Or: going down the rabbit hole. What you initially looked for, might not be what you ended up finding.

If you look up the definition for *digital humanities*, you might find many. In the period of 2009-2014 the project Day of DH collected 817 definitions (Gold, 2012). Many of these mention digital tools and methods. Often these tools and methods are applied to big datasets in the form of text or sometimes coordinates, for visualizing or for analyzing. What is often overlooked is how we work with the material and make for the small, and sometimes accidental, findings. To fully understand what digitizing does to an archive, we need to include the ways of how the human mind works and makes connections. Digitizing is not solely a technical process, but an undertaking that produces knowledge.

1. References

Berg-Nordlie, Mikkel. 2019. «Sáhkku.» Sáhkku | Reaidu.

Berg-Nordlie, Mikkel & Ida Tolgensbakk 2024. Hva med å lære et eldgammelt samisk brettspill? – SAMLA

Foldvik, Therese. In press 2026. Digital nærlesing – fra gotisk håndskrift til digital tekst. Tidsskrift for kulturforskning.

Gold, M. K. (Red.). (2012). *Debates in the digital humanities* (1st ed.). University of Minnesota Press.

Samla. What is SAMLA? - Samla

References

Berg-Nordlie, Mikkel. 2019. «Sáhkku.» Sáhkku | Reaidu.

Berg-Nordlie, Mikkel & Ida Tolgensbakk 2024. Hva med å lære et eldgammelt samisk brettspill? – SAMLA

Foldvik, Therese. In press 2026. Digital nærlesing – fra gotisk håndskrift til digital tekst. Tidsskrift for kulturforskning.

13:50–14:10 *SHORT PAPER* **[58]**

Appraising Archival Icebergs: Digitizing the Archive of the Fogelstad College for Women's Political and Civic Rights and Duties

Christa Shusko

KvinnSam, University of Gothenburg Library, Sweden

Keywords: *archives, digitization, feminism, history*

Cultural heritage materials have, in recent decades, been digitized in enormous amounts, to such an extent that many may assume that nearly everything housed in libraries and archives has already been digitized. Yet despite this often-overwhelming abundance of digitized cultural heritage, vast amounts of cultural heritage materials—and perhaps especially archival materials—have not and likely will never be digitized. Ian Milligan has called these materials "the great undigitized" (2022). Many scholars have noted that the selection criteria used for cultural heritage collection development as well as digitization are not neutral and may continue to perpetuate inequities within the cultural record (see, for example, D'Ignazione & Klein, 2020; Manžuch, 2017; Posner, 2016). For a variety of reasons, printed materials like books, newspapers, and periodicals have proven much easier to (mass) digitize. Archival materials may serve to challenge the homogeneity of existing digitized corpora, but they pose a number of material and epistemological challenges for digitization.

While it may be rather outrageous to argue that we should add *more* to the digital abundance that already threatens to overwhelm scholars, this paper suggests the importance of appraising—that is selecting—materials for digitization that can challenge and enrich the current digital corpus. Taking as a case study the digitization of the archive Fogelstad College for Women's Political and Civic Rights and Duties (sv: Kvinnliga Medborgarskolan vid Fogelstad), this paper argues that such archives can serve to make visible—if decidedly not digital—the persistent and vital depths of the additional archival materials that lie beneath these singular archives. In leveraging a feminist ethics of care alongside the models of slow or critical digitization (respectively Prescott and Hughes, 2018; Dahlström, Hansson, and Kjellman, 2012), it hopes to provide a model for appraising archival materials, not as an argument to digitize everything, but as an argument to more consciously select—and carefully contextualize—archival materials for digitization in ways that enrich and diversify the cultural record while respecting the vast undigitized.

The Fogelstad College for Women's Political and Civic Rights and Duties (for ease, henceforth referred to simply as Fogelstad) offered classes between 1925 and 1954, having been founded by a group of five women active in the women's suffrage movement: politician and estate owner Elisabeth Tamm, politician and factory inspector Kerstin Hesselgren, author Elin Wägner, medical doctor Ada Nilsson, and the school's principal and teacher Honorine Hermelin. Following women winning the right to vote in Sweden, these five women along with a range of like-minded supporters founded the school as a way to educate women from all classes in their new civic duties, enabling and encouraging them to take up public and professional roles for the betterment of society.

The Fogelstad (B7a) archive is housed in the archival collections of KvinnSam, the Swedish national library for gender research and a university-wide research infrastructure at the University of Gothenburg Library. Having been established in the 1950s—though its name, staffing, and funding have fluctuated greatly over the years—KvinnSam's archival holdings now include over 300 archival collections, and measure more than 600 shelf meters. The collections have especial strengths in Swedish women's history. KvinnSam undertaken some significant digitization projects in the past 20 years, most notably including the digitization of early Swedish women's periodicals as well as digitizing around 800 photographs for a photographic database (now defunct, though materials were migrated into another system). Yet archival materials have resisted digitization, and it is only in the last year—and with external funding from Riksbankens Jubileumsfond—that KvinnSam was able to undertake the digitization of one of its 311 archives. With the Fogelstad archive a relatively modest size, consisting of about 3 shelf meters, this means that less than half of one percent of KvinnSam's archival holdings have now been digitized.

Digitized materials in the Fogelstad archive such as course notes, lectures, applications, accounting books, and even the card catalogs of the school's now non-existent library reveal a network of women and organizations that exist—largely outside of the digitized record—within the holdings of KvinnSam and other archival institutions. Digitizing the entirety of the archive means that gaps within the archive can be somewhat assuaged through making connections between different categories of materials within the archive. This type of careful digitization also allows for easier identification of complementary materials in other archives. Identifying and contextualizing the networks of course participants, teachers, staff, and visitors can help illuminate the vastness of undigitized archival holdings and can serve as vital entry points for research into under-researched areas. Developing a map of KvinnSam's archival holdings in relation to the digitized Fogelstad archive can thus serve to illustrate the ways in which undigitized materials remain vital resources even as they lie beyond the digital.

References

Dahlström, M., Hansson, J., & Kjellman, U. (2012). 'As we may digitize': institutions and documents reconfigured. Liber quarterly: the journal of European research libraries. 21 (3-4): 455-474. https://doi.org/10.18352/lq.8036

D'Ignazione, C. and Klein, L. (2020). Data Feminism. Cambridge, MA: MIT Press. https://data-feminism.mitpress.mit.edu

Kvinnliga Medborgarskolan vid Fogelstad arkiv. B7a. KvinnSam, University of Gothenburg Libraries. Digitized archive available here: http://urn.kb.se/resolve?urn=urn:nbn:se:alvin:portal:record-115435

KvinnSam. (2025) Archival Collections. https://kvinnsam.ub.gu.se/en/archival-collections

Manžuch, Z. (2017). Ethical issues in digitization of cultural heritage. Journal of Contemporary Archival Studies 4(4):1-17. https://elischolar.library.yale.edu/jcas/vol4/iss2/4

Milligan, I. (2022). The Transformation of Historical Research in the Digital Age. Cambridge: Cambridge University Press. DOI: https://doi.org/10.1017/9781009026055

Posner, M. (2016). What's Next: The Radical Unrealized Potential of Digital Humanities. In Gold, M. and Klein, L. (Eds.). Debates in Digital Humanities 2016. Minneapolis: University of Minnesota Press. https://doi.org/10.5749/j.ctt1cn6thb.6

Prescott, A. & Hughes, L. (2018). Why Do We Digitize? The Case for Slow Digitization. Archive Journal. Why Do We Digitize? The Case for Slow Digitization - Archive Journal

14:10–14:30 *SHORT PAPER* **[59]**

Fragments, bias, lost voices and measuring silence in the digitised archive: the case of early modern maritime disability

Catherine Beck

Lunds Universitet, Sweden

Keywords: *History, Bias, Disability, Transcription, Search*

The digital technologies that shape our present also shape our past. Rapid advances in archival digitisation promise researchers and the public unprecedented access to manuscripts which claim to expose histories from below across archives at a global scale. However, historians of marginalised and hidden histories have also warned how the 'data abundance' of digitised archives encourages positivist approaches that obscure histories fragmented across archives and accessed by reading sources 'against the grain' and interpreting archival silences.[1] The way we digitally store, retrieve and analyse historical data can marginalise these vital intersectional histories of gender, race, sexuality and disability – casting them adrift on currents of infrastructure design, modern algorithmic or terminological biases and under-theorised use of computational methods in historical research.

This short paper introduces the new project FLOTSAM (Marie Skłodowska-Curie Actions, Horizon Europe, 09/2025-08/2027 ID: 101202270), which investigates how the way we structure, store and retrieve the wealth of newly available historical data in digitised early modern manuscript archives impacts these hidden histories. It asks how we can navigate not only the messiness of early modern historical data (including data fragmentation and Hand-Written Text Recognition errors)[2] but also the biases that were originally inscribed in the sources, which can be amplified, or new biases imposed, through our use/adaptation of knowledge organisation systems and language models.[3] Drawing on my domain expertise, FLOTSAM uses early modern maritime disability, broadly defined within a model

of *bodymind***[4]** impairment and difference, as a test case for the impact of digital methods on marginalised and hidden histories. The relationality and particularly the conceptual fluidity of early modern 'disability' makes it a challenging but important test case to expose the limits and potential solutions for bias and/or rigidity of digital methods and interoperable structured data.[5] FLOTSAM's ultimate aim is to add to the growing work promoting the ethical use of digital methods and safeguarding marginalised perspectives (e.g. see Combatting Bias Toolkit)[6] but also to explore how individual researchers, as we so often are in the field of historical research, can ethically adapt methods/tools (such as BERT modelling, Named Entity Recognition)[7] and contribute our specialist knowledge as interoperable data without amplifying or imposing new harmful biases.[8]

This short paper shares the earliest phase of the project, namely, the experiment design and preliminary results for testing the retrievability of instances of impairment and difference in both a dedicated corpus (which I am currently constructing from the Swedish Seamens' Houses and naval pension records) and pre-existing digitised and automatically transcribed collections in which disability appears sporadically, or which we traditionally access by reading sources against the grain. The experiment design initially focuses on two main aspects: 1) testing the relationship between HTR errors and the early modern language of disability[9]; 2) search efficacy and expansion, including downstream impact of HTR errors[10] and the suitability of e.g. topic modelling for corpus creation,[11] for developing a more complex search(es) to tackle archival silences.[12] This short paper will present the design, challenges, and the preliminary results for one or both aspects depending on the progress I manage to make by the time of revised abstract/full paper submission. I also welcome constructive feedback on my approach.

References

[1] K.M. Hunter 'Silence in Noisy Archives: Reflections on Judith Allen's 'Evidence and Silence – Feminism and the Limits of History' (1986) in the Era of Mass Digitisation' Australian Feminist Studies 32:91-2 (2017): 202–12; M. Manoff 'Mapping Archival Silence: Technology and the Historical Record' in Foscarini et al (eds) Engaging with Records and Archives: histories and theories (Facet Publishing 2016), 65; L. Putnam, 'The Transnational and the Text-Searchable. Digitized Sources and the Shadows They Cast' The American Historical Review 121:2 (2016): 377-402; M. Goebel 'Ghostly Helpmate: Digitization and Global History' Geschichte und Gesellschaft, 47:1 (2021): 35-57.

[2] H. Huistra & B. Mellink 'Phrasing history: Selecting sources in digital repositories' Historical Methods 49:4 (2016): 220-229; S. Lässig 'Digital History: Challenges and Opportunities for the Profession' Geschichte und Gesellschaft 47:1 (2021): 5–34.

[3] R. Risam New Digital Worlds: Postcolonial Digital Humanities in Theory, Praxis and Pedagogy (Northwestern University Press, 2019); C. D'Ignazio & L.F. Klein, Data Feminism, (MIT Press, 2020); M.N. Smith 'Frozen Social Relations and Time for a Thaw: visibility, exclusions and considerations for postcolonial digital archives' Journal of Victorian Culture, 19:3, (2014): 403-10; E. Manjavacas, L. Fonteyn. 'Adapting vs. Pre-training Language Models for Historical Languages'. Journal of Data Mining and Digital Humanities, 2022, NLP4DH,10.46298/jdmdh.9152.

[4] M. Price, 'The bodymind problem and the possibilities of pain', Hypatia, 30:1 (2015): 268–84.

[5] S. White, 'Crippling the Archives: Negotiating Notions of Disability in Appraisal and Arrangement and Description', Society of American Archivists 75:1 (2012): 109–124; Richards & Burch, 165; P. Horn & B. Frohne 'On the fluidity of 'disability' in Medieval and Early Modern societies. Opportunities and strategies in a new field of research' in Barsh, et al (eds) The Imperfect Historian: Disability Histories in Europe (Peter Lang, 2013); J. Gebke & J. Heinemann, 'Dealing with definitional voids: DisAbility in Early Modern Europe', Frühneuzeit-Info, 21 (2020): 8.

[6] O.M Mastromichalakis, J. Liartis, K. Rose, A. Isaac, G. Stamou 'Don't Erase, Inform! Detecting and Contextualizing Harmful Language in Cultural Heritage Collections' arXiv preprint (2025); L. Havens, M. Terras, B. Bach, B. Alex, 'Uncertainty and Inclusivity in Gender Bias Annotation: An Annotation Taxonomy and Annotated Datasets of British English Text'. in Proceedings of the 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP) (2022): 30–57, Seattle, Washington. Association for Computational Linguistics; L. Havens, B. Alex, & M. Terras, 'Confronting Gender Biases in Heritage Catalogues: A Natural Language Processing Approach to Revisiting Descriptive Metadata' in The Routledge Handbook of Heritage and Gender (2025) 351-366.

[7] M. Luthra et al 'Unsilencing Colonial Archives via Automated Entity Recognition', Journal of Documentation (2023): 1-24; M. van Erp et al. 'Slicing and Dicing a Newspaper Corpus for Historical Ecology Research' Knowledge Engineering and Knowledge Management (2017): 470-484; G. Kuys, & A. Scherp 'Representing Persons and Objects in Complex Historical Events using the Event Model F', Journal of Open Humanities Data, 8:0 (2022): 22; S. Verkijk & P. Vossen 'Sunken Ships Shan't Sail: Ontology Design for Reconstructing Events in the Dutch East India Company Archives' CHR 2023:320-332; M. Ehrmann, et al. 'Named Entity Recognition and Classification in Historical Documents: A Survey.' ACM Comput. Surv. 56:2, 27 (2024); L.F. Klein, 'The image of absence: archival silence, data visualization, and James Hemings.' American Literature, 85:4 (2013): 661–88; A. Ortolja-Baird & J. Nyhan 'Encoding the haunting of

an object catalogue: on the potential of digital technologies to perpetuate or subvert the silence and bias of the early-modern archive' Digital Scholarship in the Humanities, 37:3 (2022): 844–867; H.S.Jensen 'Digital Archival Literacy for (All) Historians.' Media History, 27:2, (2020): 251–265; L. Jaillant & K. Aske 'Are Users of Digital Archives Ready for the AI Era? Obstacles to the Application of Computational Research Methods and New Opportunities' J. on Comp. & Cult. Heritage 16:4, 87 (2024): 1-16.

[8] G. Brilmyer, 'Archival assemblages: applying disability studies' political/relational model to archival description.' Archival Science 18 (2018): 95–118 https://doi.org/10.1007/s10502-018-9287-6

[9] M. Barget, K. Hufkens, 'Best practices in pre- and post-ATR for historical research' at DARIAH-EU ANNUAL EVENT 2025 The Past (dae2025), Göttingen, Germany. Zenodo. https://doi.org/10.5281/zenodo.15703825; J. Burchardt, 'Source criticism, bias, and representativeness in the digital age: A case study of digitized newspaper archives.' Digital Humanities in the Nordic and Baltic Countries Publications. 6:1 (Sep. 2024) https://doi.org/10.5617/dhnbpub.11512

[10] S. Krivulskaya 'The Crimes of Preachers: Religion, Scandal, and the Trouble with Digitised Archives' Digitised Newspapers – A New Eldorado for Historians?: Reflections on Tools, Methods and Epistemology, eds. Estelle Bunout, Maud Ehrmann and Frédéric Clavert, (Berlin, Boston: De Gruyter Oldenbourg, 2023) 379-394. https://doi.org/10.1515/9783110729214-017 ; S. Mutuvi, A. Doucet, M. Odeo, A. Jatowt, 'Evaluating the Impact of OCR Errors on Topic Modeling.' In: Dobreva, M., Hinze, A., Žumer, M. (eds) Maturity and Innovation in Digital Libraries. ICADL 2018. Lecture Notes in Computer Science (2018), vol 11279. https://doi.org/10.1007/978-3-030-04257-8_1

[11] J. Wilson Black, 'Creating specialized corpora from digitized historical newspaper archives: An iterative bootstrapping approach', Digital Scholarship in the Humanities, 38:2, (June 2023): 779–797, https://doi.org/10.1093/llc/fqac079

[12] S. Singh, H. Suleman, 'A Comparison of Information Retrieval Pre-processing Algorithms Applied to African Historical Data'. In: Tseng, YH., Katsurai, M., Nguyen, H.N. (eds) From Born-Physical to Born-Virtual: Augmenting Intelligence in Digital Libraries. ICADL 2022. Lecture Notes in Computer Science, vol 13636. (Springer, Cham: 2022) https://doi.org/10.1007/978-3-031-21756-2_18 ; J. Guldi, 'Critical Search: A Procedure for Guided Reading in Large-Scale Textual Corpora' Journal of Cultural Analytics (December 2018); D.C.S. Wilson, M.C. Ardanuy, K. Beelen, B. McGillivray, R. Ahnert, 'The Living Machine: A Computational Approach to the Nineteenth-Century Language of Technology' Technology and Culture 64:3 (2023): 875-902; J.W. Rettberg, 'Algorithmic failure as a humanities methodology: Machine learning's mispredictions identify rich cases for qualitative analysis'. Big Data & Society, 9(2) (2022). https://doi.org/10.1177/20539517221131290; M. Barget, S. Schreibman 'Feminist DH: A Historical Perspective: Excavating the Lives of Women of the Past.' Feminist Digital Humanities: Intersections in Practice, eds. S. SCHREIBMAN and L. M. RHODY (University of Illinois Press, 2025) 35–58; S Ames, L. Havens, 'Exploring National Library of Scotland datasets with Jupyter Notebooks.' IFLA Journal, 48:1(2022), 50-56. https://doi.org/10.1177/03400352211065484 ; R. Pierce, S. Humlesjö. 'Ghost in the Archives? The Search for Feminist and Queer Archival Materials in Sweden' Digital Humanities in the Nordic and Baltic Countries Publications 7:3 (2025): https://doi.org/10.5617/dhnbpub.12258 ; A Järvelin, H. Keskustalo, E. Sormunen, M. Saastamoinen, K. Kettunen, 'Information retrieval from historical newspaper collections in highly inflectional languages: A query expansion approach' Journal of the Association for Information Science and Technology 67:12 (2015) 2928-2946 https://doi.org/10.1002/asi.23379 ; S. Jentzsch, C. Turan, 'Gender Bias in BERT - Measuring and Analysing Biases through Sentiment Rating in a Realistic Downstream Classification Task'. In Proceedings of the 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP), (2022) 184–199, Seattle, Washington.

14:30–14:50 *SHORT PAPER* **[60]**

The urge to digitise: Medieval materials in museum and archival collections

Olga Zabalueva[1], Polina Ignatova[2]

[1] *Umeå university, Sweden*

[2] *Independent scholar*

Keywords: *digital heritage, digitisation of museum collections, medieval manuscripts, knowledge circulation*

The proposed paper is going to investigate the digitalisation, digitisation, and dissemination of medieval manuscripts by archives and museums. When it comes to narrating and popularising history, the texts originating from the Middle Ages (c.500 - c. 1500) are a double-edged sword. On the one hand, the weird and funny medieval manuscript illuminations (illustrations in medieval manuscripts) have become popular among general audiences. One can find plenty of them on Pinterest, there are accounts on X dedicated to sharing fragments from medieval manuscripts, and Facebook communities, which convert manuscript illuminations into memes. There are also video games featuring characters, based on medieval manuscript illuminations, such as Illuminati and Pentiment. However, despite the increased

visibility of medieval sources, their original point and purpose remain obscure for non-specialist audiences, reinforcing false assumptions about the Middle Ages. In this paper, we will address first, the ways in which museums and archives as the institutions which in popular beliefs supposed to produce authentic and truthful knowledge, employ digitalisation and digitisation to share the history of medieval sources. Second, we will highlight both opportunities and challenges that digitalisation and digitisation produce for scholars. For example, while online manuscript collections make medieval history a more accessible discipline, digitisation is an expensive process. Consequently, the ideologically significant (e.g. the Magna Carta) or more aesthetically pleasing sources get digitised first, meaning that less appealing manuscripts risk becoming erased from public history. Whereas AI can potentially speed up the process of setting up online or physical exhibitions, AI-generated materials risk containing mistakes, which can create false historical knowledge (e.g. confusing AI-generated annotations can be found in the Swedish History Museum). The digitisation of museum collections, among other processes and outcomes, includes placing images and metadata about the museum objects into diverse digital databases. There are also local digital solutions in museums (Golub et al. 2022) and international platforms, such as the European digital library Europeana (Manikowska 2019). As Golub, Kamal and Vekselius point out, cultural heritage metadata (representations of information objects) is often a direct transition from paper catalogue records to online catalogues (2021). Online databases require a complex process of translating the categories used by individual institutions to inscribe their collections into a universal digital heritage vocabulary. This translation has proven to be more complex in museums than in libraries and archives due to the variety of collected materials (Wittgren 2013). The medieval manuscripts in museum collections can often be listed as “objects”, adding the dimension of materiality to the text, which is much harder to digitise in the current state of technological development. Dahlström, Hansson and Kjellman (2012) differentiate between “mass digitisation” and “critical digitisation”. The former is characterised by a large amount of data, a systemic approach, and an omission of intellectual, interpretative, and qualitative aspects, while the latter involves making choices, selecting, omitting, and interpreting data. Wittgren (2013) studied the “knowledge cultures” in museums and how the format and the structure of museum catalogues affect the processes of integrated digitisation. In this paper we aim to investigate how an entirely new type of medium - the digital realm - transforms the ways of recording, storing, and accessing knowledge, specifically about the history of Middle Ages. It is the starting point for the research project which focuses on the processes of curating of which types of knowledge become preserved, and which perish, and why. We also study what societal consequences can the mass transition to a new medium bear in a historical perspective, when both new and old forms of medium are existing concurrently.

References

Dahlström, Mats, Hansson, Joacim, and Kjellman, Ulrika. 2012. “As we may digitize” - institutions and documents reconfigured. LIBER Quarterly, 21(3–4), 455–474. https://doi.org/10.18352/lq.8036

Golub, Koraljka, Kamal, Ahmad M., and Vekselius, Johan. 2021. Knowledge organisation for digital humanities. In Information and Knowledge Organisation in Digital Humanities, edited by Koraljka Golub and Ying Hsang Liu, 1–22. Routledge. https://doi.org/10.4324/9781003131816-1

Golub, Koraljka, Ziolkowski, Pawel Michal, and Zlodi, Goran. 2022. Organizing subject access to cultural heritage in Swedish online museums. Journal of Documentation, 78(7), 211–247. https://doi.org/10.1108/JD-05-2021-0094

Manikowska, Eva. 2019. Cultural Heritage in the European Union. In Cultural Heritage in the European Union: A Critical Inquiry into Law and Policy, edited by Andrzej Jakubowski, Kristin Hausler, and Francesca Fiorentini, 417–444. https://doi.org/10.1163/9789004365346

Wittgren, Bengt. 2013. Katalogen - nyckeln till museernas kunskap? Om dokumentation och kunskapskultur i museer. Umeå University Press.

14:50–15:20 *LONG PAPER* **[61]**

A Best-Practice Pipeline for Nuanced, Reproducible Bibliographic Data Science -- Case VD17

Eetu Mäkelä[1], Thea Lindquist[2]

[1] *University of Helsinki, Finland*

[2] *University of Colorado Boulder, United States*

Keywords: *bibliographic data science, data enrichment workflows, reproducibility*

In our collaboration between historians and computer scientists, we have been exploring how the German VD17 (Das Verzeichnis der im deutschen Sprachraum erschienenen Drucke des 17. Jahrhunderts) retrospective bibliographic database could be used as a data source to answer questions about the dynamics of the Fruchtbringende Gesellschaft (1617–1680), or Fruitbearing Society, the first and largest academy in early modern central Europe (Ball 2008).

The Society exercised considerable influence on the development of thought, culture, and identity in early modern Germany and has long been the subject of scholarly interest. It existed for over sixty years during a time of significant upheaval in the Holy Roman Empire, a decentralised patchwork of over 350 territories. In this unstable situation, the Society aimed to “create a cultural nation based on sociability and patriotism” (Ball and van Dixhoorn 2023, 244) by championing an agenda consisting of two major goals: first, the cultivation of virtue, and second, the development of the German language.

Print was an effective means for members to disseminate and to invite engagement with these ideas. These publications were also a location of intellectual and social discourse among members, who were dispersed across central Europe and beyond. Thus, the application of bibliographic data science methods (Lahti et al. 2019) to study how the society members progressed its agenda through their publications and the community dynamic within the society and the intellectual and social sphere at the time are historically interesting ones.

However, with the society having “only” 890 members and, depending on how it is counted, between 1,500 and 2,000 publications relevant to this inquiry, any numeric inquiries will be much more susceptible to the ever-present data errors, omissions and skews in bibliographic data collections (Mäkelä et al. 2020, Tolonen et al. 2022, Mäkelä et al. 2025), as compared to typical large-scale pattern-finding inquiries that make use of tens or hundreds of thousands of records to diminish the amount of noise (e.g. Marjanen et al. 2025, Tolonen et al. 2018). Accuracy and flexibility, for instance for data enhancement and retention of information uncovered in bibliographic research undertaken as a part of the process, are thus important features for the pipeline for this use case.

To counter this problem, in the project, we have created a sophisticated pipeline that allows and integrates all of:

Taking in source data updates of the VD17 or its authority data

Applying computational enrichments to the source data

Applying manual corrective patches to the source data before computational enrichment

Applying manual corrective patches to override the computational enrichment

Being able to highlight errors in the original data and to extract corrections to be fed back upstream to the original curators

Further, the pipeline has been developed to be reproducible and deterministic given particular inputs, thus allowing the resulting data set to be constantly updated as any of its inputs change. Moreover, the pipeline has been integrated with diagnostic facilities for a) verifying the accurate application of manual patches, both initially as well as when the underlying data changes, and b) tracking and verifying how the end result changes when any updates are made to the data or processing rules, to verify that work all the time moves forward, instead of back.

In this presentation, we will go over this pipeline, describing in detail how the different data sources and transformations in it are implemented and made interoperable, how our diagnostic processes increase trust in the accuracy of the final output, and how much of a difference even small corrections make for analytic findings.

References

Ball, Gabriele. “Alles zu Nutzen – The Fruchtbringende Gesellschaft (1617–1680) as a German Renaissance Academy.” In The Reach of the Republic of Letters: Literary and Learned Societies in Late Medieval and Early Modern Europe, edited by Arjan van Dixhoorn and Susie Speakman Sutch, 389–422. Leiden; Boston: Brill, 2008. https://doi.org/10.1163/ej.9789004169555.i-522.87.

Ball, Gabriele, and Arjan van Dixhoorn. “Transformations: The Rise of New Institutions.” In Performative Literary Culture: Literary Associations and the World of Learning, 1200–1700, edited by Arjan van Dixhoorn and Susie Speakman Sutch, 171–200. Brill’s Studies in Intellectual History, vol. 347. Leiden; Boston: Brill, 2023. https://doi.org/10.1163/9789004546196.

Das Verzeichnis der im deutschen Sprachraum erschienenen Drucke des 17. Jahrhunderts (VD17). “Start – VD17.” Accessed 15 July 2025. http://www.vd17.de/.

Marjanen, Jani, Tuuli Tahko, Leo Lahti, and Mikko Tolonen. “Book Printing in Latin and Vernacular Languages in Northern Europe, 1500–1800”. In Hanssen JM, Furuseth S, editors, The Hermeneutics of Bibliographic Data and Cultural Metadata. Oslo: National Library of Norway. 2025. p. 27-66.

Mäkelä Eetu, Krista Lagus, Leo Lahti, Tanja Säily, Mikko Tolonen, Mika Hämäläinen, Samuli Kaislaniemi, and Terttu Nevalainen. “Wrangling with non-standard data.” In S. Reinsone, I. Skadiņa, A. Baklāne, & J. Daugavietis (Eds.), Proceedings of the Digital Humanities in the Nordic Countries 5th Conference: Riga, Latvia, October 21-23, 2020 (pp. 81-96). (CEUR Workshop Proceedings; No. 2612). CEUR-WS.org. http://ceur-ws.org/Vol-2612/paper6.pdf

Mäkelä, Eetu, Thea Lindquist, Narges Azizifard, and Julius Arnold. “Fruchtbringende Gesellschaft Member Publication Patterns in the VD17.” Digital Humanities in the Nordic and Baltic Countries Publications 6, 1 (2024). https://doi.org/10.5617/dhnbpub.11485.

Mäkelä, Eetu, James Misson, Devani Singh, Mikko Tolonen. “Opening the Black Box of EEBO”. Digital Scholarship in the Humanities. To appear (2025). https://doi.org/10.1093/llc/fqaf086

Tolonen, Mikko, Eetu Mäkelä, and Leo Lahti. "The Anatomy of Eighteenth Century Collections Online (ECCO)." Eighteenth-Century Studies 56, no. 1 (2022): 95-123. https://dx.doi.org/10.1353/ecs.2022.0060.

Lahti, Leo, Jani Marjanen, Hege Roivainen, and Mikko Tolonen. 2019. “Bibliographic Data Science and the History of the Book (c. 1500–1800).” Cataloging & Classification Quarterly 57 (1): 5–23. doi:10.1080/01639374.2018.1543747.

Tolonen, Mikko, Leo Lahti, Hege Roivainen, and Jani Marjanen. 2018. “A Quantitative Approach to Book-Printing in Sweden and Finland, 1640–1828.” Historical Methods: A Journal of Quantitative and Interdisciplinary History 52 (1): 57–78. doi:10.1080/01615440.2018.1526657.

Session 3C — 13:30–15:10

13:30–14:00 **Non-Canonical Use of a Historical Dictionary: User Experience and User Data**
Sheryl McDonald, Tarrin Wills

14:00–14:20 **Icelandic eponyms: Detection and Representation in Lexicographic Sources**
Ellert Johannsson, Steinthor Steingrimsson

14:20–14:40 **The effect of translatorial signals in the source language recognition**
Manex Agirrezabal, Seán Vrieland, Iliana Kandzha

14:40–15:10 **Неком Дрина тече десно, неком Дрина лијево тече (For some, the Drina flows to the right, for others, the Drina flows to the left)**
Sasha Rudan Kelbert, Lazar Kovacevic, Sinisa Rudan

13:30–14:00 *LONG PAPER* **[62]**

Non-Canonical Use of a Historical Dictionary: User Experience and User Data

Sheryl McDonald, Tarrin Wills
University of Copenahgen, Denmark

Keywords: *lexicography, user interfaces, user experience, digital editing, Old Norse*

1. Introduction

The Dictionary of Old Norse Prose (ONP) began in 1939 as a project supplementing existing historical dictionaries’ lacking coverage of non-canonical Old Norse texts. It has since developed into a full analysis of the Old Norse lexicon in prose works, but retains its extensive analysis of new, unusual and otherwise unconventional texts in Old Norse.

Historical dictionaries are traditionally based on extensive reading and excerption of a (print) corpus, selecting words for analysis that are copied along with their context and references to the source. The focus of ONP in its first decades was precisely this. The resulting card index, now digitized, contains excerpts for around 7% of the approximately 11 million words in the underlying (undigitized) corpus (Wills and Johannsson 2019: 119). This means that almost all unusual words, usages and phrases in

any given Old Norse prose text have been or will be analyzed and defined by ONP, which is now approximately two thirds complete.

ONP works only from manuscript evidence and therefore uses editions that closely reference the underlying manuscripts. The Old Norse textual corpus is very large and only a small proportion has been edited in digital form; of this, an even smaller proportion of digital editions are sufficiently close to the original manuscripts to be used as the basis for ONP's analysis of the lexicon.

ONP has therefore a very extensive collection of carefully curated data concerning analysis of the language of all Old Norse texts, including non-canonical and otherwise difficult and inaccessible works. This paper addresses the problem of how a conventional dictionary can be used in unconventional ways to give users access to and engage them with these works.

We address the following questions:

1. How can a dictionary framework be developed to give users access to the underlying corpus text and its analysis?
2. How can the user experience be enhanced to engage users with the texts, particularly the non-canonical materials that form part of the dictionary's corpus?

2. ONP Reader

The digitization process of 2005–10 included scanning almost all printed editions from which ONP's citations are drawn. The web publication of ONP since 2011 has included these scanned images in certain contexts: dictionary entries (both edited and unedited) include a list of all citations for each word, with links to further information about each citation, including the page and line number of the edition from which it is taken. This includes in most cases also the corresponding page of the edition from which the citation was drawn. These images are subject to an agreement with the Danish copyright agency (CopyDan) so that more recent editions can be reproduced in this way.

More recently, it became apparent that this detailed information connecting citations and their analysis with the scanned editions from which they are drawn could be used in a novel way: to generate glosses of words from editions. This could then be presented alongside the scanned images of the editions. The resulting facility was named "ONP Reader" (Wills and Johannsson 2019). It effectively provides a gloss to the editions ONP uses. The detail of excerption and analysis means that a very large proportion of difficult (e.g. low frequency) words in any given text are analyzed and made available to users in this way.

ONP is, however, still a dictionary. Users normally expect to begin using an online dictionary with a word search rather than by looking for a text. The "reader" view is therefore deep within the website: users must find a word, click on a citation and then on the reader view link to see the glossed page from which the citation is taken. In user experience (UX) terms, the interaction cost for accessing the ONP Reader is very high.

3. User experience of the ONP Reader

UX4DH – User Experience for Digital Humanities is a resource for UX design in digital humanities. Its guide to user research (UXR) outlines four methods, only two of which are currently within the resource restrictions of ONP: user interviews and analytics review. Some of the observations below under qualitative assessment are based on informal interviews with users, who share their experiences at conferences and meetings, or who contact or cite ONP directly. Although GDPR restrictions mean that ONP cannot use Google Analytics — the method outlined by UX4DH to analyze user behavior — we present the quantitative insights gained from custom web analytics collection from ONP's web application.

3.1. Qualitative

One of the challenges faced when considering how to improve the user experience for ONP is the knowledge of the existence of a range of different types of users accessing ONP's resources, coupled with a lack of information about these users. This is a common problem among digital humanities projects that have been designed first and foremost to showcase research results, without due consideration for users or human-centered design principles (Miller 2024: 70–71). While we have yet to carry out targeted user research to collect more detailed data that would further define ONP's userbase through a set of clearly defined personas (e.g. Laitan 2021; Miller 2024), qualitative evidence

gained through informal discussions with colleagues and students in the field of Old Norse Studies suggests the presence of at least two different users with different levels of experience working with Old Norse, and different levels of intuition to draw on when navigating ONP's website and resources. We used workshops and conferences to gain feedback from users in group and individual settings including the International Saga Conference (Katowice), Summer School in Scandinavian Manuscript Studies (Reykjavík) and invited papers and teaching in Prague, Trento, Verona and Zürich.

At one end of the spectrum is the advanced user: an expert in Old Norse. This advanced user is the type envisioned for ONP from its inception and the dictionary and its auxiliary resources have been built with them in mind. However, another user type also exists: the beginner, with limited experience with Old Norse. Beginners have moreover little to no experience reading unnormalized Old Norse in manuscripts and diplomatic editions. The presentation of data in ONP in its current state, therefore, is highly inaccessible for beginners, while catering to advanced users.

Benefits of the current user interface include a direct link from the homepage to a list of editions from which ONP Reader can be accessed, giving digital access to editions. However, navigating this requires advanced knowledge of the text one wishes to access, as multiple editions and manuscripts are often represented under each text. For beginners, the action of accessing a text in ONP Reader to aid comprehension of the original language becomes complex.

To increase accessibility for beginners without over-simplifying or limiting functionality for experts, we believe a solution to be implementing small changes improving the presentation of data contained in the lists of ONP Reader's editions and manuscripts. For example, by indicating a "best" edition and the main or "best" manuscript in cases where multiples are listed, ONP Reader's list of editions and manuscripts could become more accessible.

3.2. Quantitative

In accordance with GDPR, ONP no longer uses Google Analytics to track user activity. Instead, user interactions are logged internally and the data managed in accordance with GDPR, institutional and governmental policies. Users who consent to tracking are assigned a random identifier, which allows local analytics to potentially follow their behavior through the site; users who do not consent have activity recorded by IP address. To save on storage space, referrer information, which would allow better tracking between pages, was not recorded until recently. At the time of writing, we have therefore only a short-term picture of how users access ONP Reader. We will present more detailed information at the DHNB conference, but the current situation can be characterized as follows.

The total number of visits logged by internal analytics has been largely consistent since these analytics began in 2022. The page views of the ONP Reader have remained in proportion to the number of views for conventional dictionary pages such as words and citations, between 14–18% of all such views in this period. Preliminary statistics suggest that very few users (0.3%) who use the Reader are accessing it via the ONP Reader list linked on the start page. While many arrive via dictionary citations, a larger proportion come to ONP Reader using links from the index and bibliography pages.

Preliminary data suggest that for users wanting to access ONP Reader texts, whether unusual or canonical, they search specifically for the texts, rather than arriving via the lexicon. This suggests that ONP is being used in a novel way: as a text repository for both canonical and non-canonical editions of Old Norse. The lack of referrals out of the Reader, which links words, citations and other resources, suggests that these pages are themselves the users' goal in using the site. This reflects perhaps that ONP has made a well-curated, high quality and enhanced corpus of Old Norse editions accessible to a broad user base, even as it is a secondary goal of a dictionary project.

References

Aldís Sigurðardóttir, Alex Speed Kjeldsen, Bent Chr. Jacobsen, Christopher Sanders, Ellert Þór Jóhannsson, Eva Rode, Helle Degnbol, James E. Knirk, Johnny Lindholm, Maria Arvidsson, Pernille Ellyton, Sheryl McDonald, Simonetta Battista, Tarrin Wills, Þorbjörg Helgadóttir, Þórdís Edda Jóhannesdóttir, eds. 1989– . Dictionary of Old Norse Prose. https://onp.ku.dk.

Laitan, Maria. 2021. "UX Methods for Digital Humanities Projects." Poster with transcript and slides presented at HighEdWeb Association Annual Conference, October 4–5 2021. https://membership.digicol.org/2021-annual-conference-resources/ux-methods-digital-humanities-project/.

Miller, A. 2024. "Inclusive Design: A Method and Craft of Transforming Digital Humanities with User Experience." In Digital Humanities in the Library, edited by Arianne Hartsell-Gundy, Laura R. Braunstein, and Liorah

Golomb. 2nd edn. Association of College and Research Libraries: 69–88. https://jewlscholar.mtsu.edu/handle/mtsu/7396.

Williams, George H. 2012. “Disability, Universal Design, and the Digital Humanities.” In Debates in the Digital Humanities, edited by Matthew K. Gold. Minneapolis: University of Minnesota Press: 202–212.

Wills, Tarrin and Ellert Þór Jóhannsson. 2019. “Reengineering an Online Historical Dictionary for Readers of Specific Texts.” In Electronic Lexicography in the 21st century: Smart Lexicography. Proceedings of the eLex 2019 Conference. 1–3 October 2019, Sintra, Portugal, edited by Iztok Kosem, Tanara Zingano Kuhn, Margarita Correia, José Pedro Ferreria, Maarten Jansen, Isabel Pereira, Jelena Kallas, Miloš Jakubíček, Simon Krek, and Carole Tiberius. Brno: Lexical Computing CZ, s.r.o.: 116–129.

UX4DH — User Experience for Digital Humanities. https://ux4dh.byu.edu.

14:00–14:20 *SHORT PAPER* **[63]**

Icelandic eponyms: Detection and Representation in Lexicographic Sources

Ellert Johannsson, Steinthor Steingrimsson

The Árni Magnússon Institute for Icelandic Studies, Iceland

Keywords: *Eponyms, Icelandic lexicography, Word formation, Onomastics*

This project investigates eponyms in Icelandic dictionaries, that is, lexical items derived from personal names, such as *grettistak* ‘a great feat of strength’ (from *Grettir Ásmundarson*, a semi-historical figure from a well-known medieval saga) or *þórðargleði* ‘schadenfreude’ (from *Þórður,* a character in a famous Icelandic literary work from the 20th century). Eponyms constitute a particularly rich and productive class of word formations, as they link linguistic creativity with cultural and historical references. Despite their visibility in everyday usage eponyms have received little systematic attention in Icelandic lexicography or in studies of Icelandic word formation. The project seeks to provide a comprehensive account of how Icelandic eponyms are formed, recorded, and represented in different lexicographic sources.

The research focuses on major dictionaries: Sigfús Blöndal’s *Íslensk-dönsk orðabók* (1920–1924), *Íslensk nútímamálsorðabók* (ÍN), and selected digital lexicographic and corpus-based resources, such as *Risamálheildin* [A very large corpus of Icelandic texts] (cf. Steingrímsson et al 2018). These sources represent distinct stages in the documentation of Icelandic vocabulary, from early twentieth-century lexicography to contemporary digital dictionaries and allow for a diachronic perspective on how eponyms enter and evolve within the lexicon. By comparing their inclusion and treatment across these resources, the study aims to reveal both lexical continuity and changes in lexicographic norms.

A central methodological challenge lies in identifying eponyms automatically, as they are not explicitly labeled or tagged as such in any existing dictionary. To address this, the project employs a data-driven approach that combines morphological and lexical information from the *Beygingarlýsing íslensks nútímamáls* (BÍN), a comprehensive database of Icelandic inflectional paradigms, with digital dictionary data. The inflectional forms of Icelandic personal names, especially genitive forms, such as *Jóns* form the name *Jón*, *Guðmundar* from *Guðmundur*, or *Helgu* from *Helga,* which would form a part of a compound, serve as search patterns for detecting potential eponymic formations in dictionary headwords. This method makes it possible to reveal eponyms formed with Icelandic personal names as a morphological element.

Once identified, candidate eponyms will be classified according to morphological structure, word-formation pattern e.g. compounding or derivational endings such as *-legur*, *-fræði*, *-ismi* and semantic types and semantic type (e.g. names of persons, both historical figures and fictional characters as well as mythological beings.). Quantitative analysis will explore the relative productivity of different name bases and affixes, while qualitative examination will focus on semantic motivation and cultural associations. The study will also look at how dictionaries define eponyms, i.e. whether they explicitly refer to the motivating person or treat the eponym as fully lexicalized and if these strategies have changed over time.

By integrating computational techniques with traditional lexicographic and onomastic analysis, this project contributes to the broader field of digital humanities and to Icelandic word-formation studies. It offers the first systematic overview of Icelandic eponyms and provides insight into how personal names become lexicalized and part of the active vocabulary of the language. More broadly, the project

illustrates how combining lexicographic data with digital methods can deepen our understanding of the relationship between language and cultural history.

References

Sources

BÍN = Beygingarlýsing íslensks nútímamáls. Kristín Bjarnadóttir (red.). The Arni Magnusson Institute for Icelandic studies. <bin.arnastofnun.is> (October 2025)

Blöndal, Sigfús. 1920-24. Íslensk-dönsk orðabók / Islandsk-dansk ordbog. Reykjavík: Prentsmiðjan Gutenberg.

ÍN = Íslensk nútímamálsorðabók. Jónsdóttir, Halldóra & Þórdís Úlfarsdóttir (red.). The Arni Magnusson Institute for Icelandic Studies. <islenskordabok.is> (October 2025).

Íslensk-dönsk orðabók [Islandsk-dansk ordbog] digtitized online <blondal.arnastofnun.is> (October 2025).

Risamálheildin (Gigaword Corpus) = RMH 2024. Stofnun Árna Magnússonar í íslenskum fræðum. <malheildir.arnastofnun.is> (september 2025)

Steingrímsson, Steinþór, Sigrún Helgadóttir, Eiríkur Rögnvaldsson, Starkaður Barkarson og Jón Guðnason. 2018. Risamálheild: A Very Large Icelandic Text Corpus. Proceedings of LREC 2018, Myazaki, Japan, 4361-4366.

14:20–14:40 *SHORT PAPER* **[64]**

The effect of translatorial signals in the source language recognition

Manex Agirrezabal, Seán Vrieland, Iliana Kandzha

University of Copenhagen, Denmark

Keywords: *source language, translation, classification, prayer*

In this work, we are developing models for the recognition of the source language in a translation. This is part of the “When Danes Prayed in German” project[1], a project funded by Independent Research Fund Denmark [1055-00040B]. The main goal of the project is to examine the role of Low German in the transition from Latin to Danish in devotional manuscripts from the late Middle Ages and the early modern era. One of the building blocks of the project will be used to examine whether texts in Danish were translated directly from Latin or via Low German. This will be done using Natural Language Processing tools. The main challenge is data scarcity. The amount of annotated data that we know that was translated from Latin/Low German is small, and is still being digitized and proofread.

In order to test source language recognition models, we are currently employing a larger corpus of texts translated from German and French to English (Lynch and Vogel, 2018). This corresponds quite well to the language families that we have in the original project. One aspect that we believe can help in the recognition of the source language are translatorial signals, which may have remained in the translations, due to specific word spellings or due to syntactic constructions (from the source language). We performed three experiments in the data by Lynch and Vogel. First of all, we ran [SV1] a baseline experiment using character bigrams and a Logistic Regression classifier. This baseline method is expanded by a list of ngram changes between language pairs. For instance, a common pair between German and English would be “sch-sh”, which can be observed in many words ending in “sh” in English (Englisch-English, fisch-fish, goulasch-gulash, …). Using Minimum Edit Distance, we build two lists of character changes (German-English and French-English) from two existing bilingual dictionaries and use it to enrich the character representations. We performed a number of experiments, with different parameters, and the main observation is that we do not manage to improve the baseline in none of the configurations. Last, but not least, we tested [SV2] whether morphosyntactic information can help in the detection of the source language. We do this by expanding the original bigram character-based representations with Part-Of-Speech patterns (bigrams and trigrams). The initial experimentation shows that results are slightly better, in terms of F1-score (0.768 vs 0.782). We further compared [SV3] the baseline (char bigrams) and the expanded representations (char bigrams+pos tags) with a few classifiers (Decision Tree, Random Forest and Linear SVM), and we observe a slight improvement in 3 out of 4 classifiers (Random Forest gets slightly worse with the expanded representations).

It seems like using character changes does not help much. This may be because of the high quality of the translations, where the translator does not perform a word-by-word translation, but a whole rephrasing of texts. Therefore, remaining sounds of the original language might be scarce. When running the experiment with POS-tags, we can observe

that POS-tag patterns are present in the top-10 relevant features (using Logistic Regression model's weights). In the future, we would like to further work on the interpretation of these results, maybe further analyzing the other classifiers (the leaves of the Decision Tree, for instance).

References

Lynch, G., & Vogel, C. (2018). The translator's visibility: Detecting translatorial fingerprints in contemporaneous parallel translations. Computer Speech & Language, 52, 79-104.

14:40–15:10 *LONG PAPER* **[65]**

Неком Дрина тече десно, неком Дрина лијево тече (For some, the Drina flows to the right, for others, the Drina flows to the left)

Sasha Rudan Kelbert[1,2], Lazar Kovacevic[3], Sinisa Rudan[2]

[1] *University of Oslo, Norway*
[2] *ChaOS, Serbia*
[3] *Inverudio, USA*

Keywords: *dialect migrations refugees translation*

This paper presents our ongoing research that addresses the **cultural heritage, history, identity, and language of Serbs in the Krajinas west of Serbia**, a population often self-identified as "[Срби] преко Дрине" (*Serbs over the Drina*). The expression itself encapsulates a complex, shifting notion of belonging, border, and identity. Formally, it refers to Serbs living west of the river Drina—beyond the political borders of present-day Serbia. Yet it simultaneously points to an internal dichotomy: from the perspective of Serbs in Bosnia, Croatia, or Dalmatia, *Serbia itself* becomes the "over-Drina" space. Historically, the Drina has functioned not only as a geographical marker but also as a **metaphorical border of cultural identity**—a threshold between the "матица" (*matica*, or homeland) and its dispersed communities.

Across centuries, this metaphorical Drina has shifted repeatedly as imperial, national, and ideological borders changed, moving Serbian communities and their cultural heritage *inside* or *outside* the matica without any physical migration at all. As Mexican Americans poignantly describe their historical experience, "We didn't cross the borders—the borders crossed us." In the case of the Serbs of the Krajinas, this experience is compounded by **actual migrations**, multiple displacements, and returns, each of which reshaped the texture of language, customs, and memory. These **entangled histories of motion and belonging** provide a uniquely fertile ground for digital humanities inquiry—an opportunity to explore the **non-canonical, peripheral, and dispersed archives** that constitute the living memory of a people whose identity has been continuously redefined by mobility and loss.

Corpus and Sources

Within this project we have begun to **collect, digitize, and annotate** a growing body of texts, images, and audio materials representing the **heritage of Serbs from the western Krajinas** and their modern diasporas. These include folk songs, travelogues, dialectal dictionaries, anthologies, and critical essays spanning from the nineteenth century to the present. Among the initial corpus are works such as:

Književnost Srpske Krajine, Dušan Ivanić

Severnom Dalmacijom: Putopisi i reportaže, Mirko Korolija

Srpske narodne pesme iz Slavonije, Đorđe Rajković

Rječnik srpskog govora Korduna, Boško M. Bućan

Glas iz Tuđine, Milka Kajganić

Antologija srpskih narodnih balada Krajine, Zdravko Krstanović

Pjesme: prevodi, obrade, Sava Mrkalj

Salo debeloga jera libo azbukoprotres, Sava Mrkalj

and early cartographic and textual sources such as *Regnum Bosniæ … cum finitimis Croatiæ, Dalmatiæ, Slavoniæ, Hung. et Serviæ partibus*.

This corpus also incorporates **contemporary testimonies** and literary works by descendants of these communities, expanding the traditional canon of Serbian literature with **voices of return, displacement, and hybrid identity**. Our aim is not merely archival but analytical: to study how **language, memory, and place co-evolve** through migration and re-contextualization.

Methodological Framework: Digital Scholarly Edition and Data Modeling

All collected materials are integrated into the **LitTerra Digital Scholarly Edition (DSE)** platform—a modular digital humanities infrastructure developed within the ColaboFlow ecosystem. The LitTerra DSE enables **multi-modal representation** of cultural data, interlinking texts, images, audio recordings, maps, and scholarly metadata through semantic graph structures. Each artifact is modeled as a **node** in a knowledge graph (KnAllEdge) enriched with descriptive, temporal, and spatial relations. This design supports *both close and distant reading* of the corpus, allowing scholars to trace connections between authors, dialectal expressions, and geographic references.

To extend accessibility, we employ **Bukvik**, a sub-platform dedicated to **linguistic processing of dialectal and archaic Serbian**. Using a combination of **natural language processing (NLP)** and **custom dialectal dictionaries**, Bukvik performs morphological and lexical mapping between historical expressions and their modern counterparts. This enables computational analysis of old texts while preserving their linguistic authenticity. For example, archaic words and localisms are automatically linked to modern Serbian equivalents, helping both readers and algorithms interpret the cultural and emotional nuances embedded in the original phrasing.

Cross-Referencing, Visualization, and Storytelling

Beyond textual analysis, we situate each item within **historical and geographic contexts**. Using **geospatial referencing**, we map the corpus onto both historical and contemporary maps, connecting places mentioned in songs, letters, and travelogues with tangible locations, people, and artifacts. By integrating visual materials—illustrations, manuscripts, portraits, and objects—we create a **"museum of words"**, a dynamic digital space where intangible cultural heritage is visualized through its tangible traces.

Importantly, we approach this not as static documentation but as **story-driven exploration**. The LitTerra interface employs **gamified narrative pathways** that invite readers to navigate cultural history as a form of interactive storytelling rather than a linear archive. Users can follow the routes of travelers, the spread of dialects, or the evolution of motifs such as exile, homecoming, or faith, effectively "reading" history through a network of interlinked cultural objects. This approach aligns with contemporary digital-heritage trends emphasizing **engagement, re-use, and emotional resonance**, moving beyond purely factual presentation toward **experiential scholarship**.

Toward Participatory Futures: CoLaboArthon

The final dimension of our project extends from documentation into **collaborative creative production**. We have developed **CoLaboArthon**, an integrated creative-writing platform connected to LitTerra. CoLaboArthon applies methods from **computational co-creation** and **collaborative storytelling** to allow writers, scholars, and community members to **compose new texts within historical, linguistic, and cultural contexts** derived from the corpus. Participants can write in the dialects of their ancestors, explore underrepresented perspectives, or reconstruct imagined voices of lost communities. The system uses contextual prompts, semantic suggestions, and geospatial references to **bridge the past and the present**, fostering what we term *"contextual re-appropriation of heritage."*

In this way, the corpus becomes not only an archive but a **living creative environment**—a place where the descendants of displaced communities can "claim" their past, experiment with identity, and participate in reconstructing collective memory. This participatory approach resonates with current debates in digital humanities about **decolonizing archives**, **empowering communities to reinterpret heritage**, and **rethinking authorship in the age of AI-augmented creativity**.

Discussion and Relevance to the DHNB 2026 Theme

Under the conference theme *Lost in Abundance – Encounters with the Non-Canonical*, our work explores the **abundance of neglected cultural materials** and the **methodological challenges** of

engaging with them. The Serbian Krajina corpus represents a **non-canonical, fragmented, and geographically dispersed body of texts**, often marginalized in national literary historiography. Digitization and computational methods enable us to **re-aggregate** these fragments into new forms of coherence without erasing their heterogeneity.

Our contribution demonstrates how **digital infrastructures can reframe peripheral archives** not as residual but as generative—producing new insights into how language, place, and identity intertwine. By combining **data modeling**, **NLP**, **geospatial mapping**, and **participatory storytelling**, we illustrate a model for digital humanities projects that move fluidly between **scholarship, public engagement, and creative expression**.

Ultimately, the project questions how cultural borders— like the metaphorical Drina—are drawn, crossed, and reimagined in digital space. It shows how digital methods can **navigate abundance without flattening difference**, allowing non-canonical voices to speak within richly interconnected, multilingual, and multimodal frameworks. In doing so, it contributes to ongoing conversations in DH about **representation, inclusivity, and the ethics of abundance** in the study of cultural heritage.

Session 4A — 16:00–18:00

16:00–16:30 **E-motion: Binary systems versus fluid identities**
Onur Kilic, Evelina Liliequist, Coppelie Cocq, Karin Danielsson

16:30–17:00 **Navigating Abundance: A Platform for AI-Augmented, Hybrid Reading of Multilingual Literary Texts**
Sasha Rudan Kelbert, Eugenia Kelbert Rudan

17:00–17:30 **Computational Phonosemantics at Scale: Measuring Sound-to-Meaning Mappings in English and Polish with Gradient Boosting and GLMs**
Szymon Pindur

17:30–18:00 **Modeling Textual Emotions in Literary Fiction**
Kirstine Nielsen Degn, Alexander Conroy, Xiaoyuan Jiang, Ali Al-Laith, Jens Bjerring-Hansen, Tanya Karoli Christensen, Ingo Zettler, Daniel Hershcovich

16:00–16:30 *LONG PAPER* **[66]**

E-motion: Binary systems versus fluid identities

Onur Kilic, Evelina Liliequist, Coppelie Cocq, Karin Danielsson
Umeå University, Sweden

Keywords: *Motion capture, queer theory, identification, binary systems, focus groups, participatory design*

Background

While the design of AI systems has shifted towards larger-scale datasets that could potentially include broader representation, critical studies on the widely used facial analysis (FA) datasets highlight persistent injustices that marginalized groups face from their interaction with AI systems. These systems often make (mis)classification of bodies depending on cisgendered, white supremacist, and ableist assumptions (Buolamwini & Gebru, 2018; Benjamin 2019; Shew 2023) and excluding those who do not fit (Scheuerman et al., 2019). In an earlier study, we discussed how these systems often rely on binary logics that can reinforce fixed and essentialist understandings of gender and sexuality, overlooking the fluidity and complexity of identities (Danielsson et al. 2023). We further argued for the need of casting a queer eye on AI systems by using queer theoretical perspectives to challenge binary systems and reflect on how these systems could be reimagined to capture fluid and complex identities (Liliequist et al. 2023). Moreover, systems that use body information such as motion and biometrics are used in policing the society and in bordering practices, raising concerns for bias (Buolamwini & Gebru, 2018) and privacy (Nair et al. 2023). Scholars also demonstrate that motion capture (mocap) systems' erasure of the markers such as race or gender during data capture creates a false assumption of neutrality, while discriminatory assumptions of computational systems persist (Chang 2019). Therefore, marginalized bodies often face two contradictive, yet recurring conditions in their interactions with AI:

invisibility, where bodies are rendered illegible or unrecognizable by the system; and hypervisibility, where bodies are surveilled and subjected to intensified control (Benjamin 2019). We take these tensions surrounding the misclassification and surveillance of bodily data as a point of departure to critically examine how AI systems are designed and trained.

In this study, we conduct explorative research using a mocap system that records bodily movements and gestures into data. We aim to challenge dominant, top-down approaches that rely on fixed, binary, and normative identity categories, arguing instead that identities are fluid, complex and context-dependent. Focusing on queer life experiences, the study explores how data for mocap can be co-created in a participatory, bottom-up dynamic, and how complex and multifaceted identities can be expressed in data.

Theoretical Framework

As much as we are concerned with what AI systems do to our bodies in the process of identification, we also ask what our bodies can do to AI systems. Drawing on posthuman feminist (Braidotti, 2022) and queer theory (Butler 1990; Puar 2020) perspectives, we approach bodies not merely as objects of surveillance and discipline but as active agents in reconfiguring and (re)imagining futures with (or despite) AI. Body, in this sense, is to be seen beyond biological/technological determinism (Braidotti 2022), and as an assemblage for identification, a journey towards becoming (Russell 2020:146). We are inspired by such theoretical understandings of the body, not as a fixed identity position, but as a speculation that transforms with technology. Our approach ties body closely with queerness, where queer is not a condition, but a horizon for imagining the otherwise (Muñoz 2009). This declaration positions bodies in openness and as futurity, highlighting its imaginative potential.

In this study, our focus is on those moments when our bodies become “data-bodies,” where we leave digital traces of our embodied presence in online spaces. The automation of bodies becomes problematic when identity-based categorization (or classification) emerges, reducing bodily traces to fixed identities. However, we can shift attention away from classification and toward the circulation of feelings. Queer scholar Jasbir Puar argues that eliminating ontological separation between matter (bodies of humans/machines) and discourse (identities) allows us to recognize the agential force of queer bodies and see identity as an “event” rather than a label (Puar, 2018; see Kilic, 2024). Inspired by such a vision, we see identities as assemblages of bodies, affects, and movements; rather than fixed labels based on language and discourse. We also see those identities based on gender, sexuality, and race as transitioning over time, changing with affective interactions, and disrupted by digital technologies (Càrdenas, 2016).

Design and Methods

Our research combines participatory design (PD) and feminist ethnography approaches that enable a bottom-up engagement with mocap data and provide tools for challenging identitarian binarism. PD is an emergent design approach for AI systems (Zytko et. al., 2022), and the examples of using PD during design of AI are still few but increasing. In this study, we are inspired by PD perspectives that focus on lived experiences of marginalized communities and takes the standpoint of the community needs (Costanza-Chock 2020). In action, with inspiration from queer theoretical perspectives, we seek to destabilize top-down, fixed, and binary data and create “queer” data sets. In doing so, we adopt an ethnographic approach to examine the technocultural situatedness of the design process. This is a crucial step, as most AI systems produce culturally situated knowledge despite their tendency to equate west centric knowledge with universal objectivity (Klipphahn-Karge et al. 2024). To this end, we employed two primary methods: focus group interviews and a mocap workshop with actors.

We designed a focus group study with sexually diverse and gender-variant participants from the LGBTQI+ community to gain insights and facilitate a participatory approach to mocap data that positions participants as co-creators rather than research subjects. Our role as facilitators, therefore, is not to direct interpretation but to create conditions for collective knowledge-making, where participants set the terms through which their affective experiences are articulated and translated into mocap data. We work with two groups: in Umeå and Malmö. This choice is not primarily about geographic diversity but about the different conditions that shape queer life and belonging in these locations. Malmö is a cosmopolitan city with an ethnically diverse population and a vibrant queer community; there, we collaborated with a queer cultural organization, which provided a community-based space and a point of engagement for recruitment and discussion. Umeå is a university city in northern Sweden with a

strong profile in queer art and cultural activity, where focus groups were held in university facilities. Participants in both cities were recruited through snowball sampling.

Each group met for two rounds following the same format. The first round functioned as co-creative design session for the planned mocap workshop. We introduced the mocap system through visual representations from our pilot recording and facilitating a discussion on body representation alongside a data labelling activity. The session concluded with collaborative storytelling, where participants co-created fictional narratives drawn from their queer life experiences, intended to be performed in the subsequent mocap workshop.

Between the two rounds of focus groups, we conducted the mocap workshop itself with actors, capturing bodily data to explore how shifts in performativity, body language, and movement can transform the production of mocap data. In this session, the collaborative narratives from focus group participants were performed. The workshop served a dual purpose: (I)generating data whose production was shaped by queer community members, (II)providing an ethnographic site for us as researchers to observe the limitations, glitches, and power structures that emerge through the use of such systems.

The second round of focus groups invited the same participants into a reflective dialogue on the data capturing process. Participants critically assessed the visual representations of the data produced, contested or affirmed the outputs generated, and speculated collectively on alternative futures for queer-AI interaction. This round centred on a "collective visualization" session, where participants took the mocap recordings as a basis for speculating on new forms of data visualization, exercising epistemic agency over how their contributions would be represented.

Following this participatory process, we plan to create data visualizations with the support of the InfraVis research infrastructure, translating the outcomes of the queer speculations generated through the focus groups and mocap workshop into their final visual form.

Discussion

A central aim in this study is to challenge binary and fixed classifications in mocap data. Therefore, we are intrigued by the ways in which fluid identities can be expressed affectively and interactively. As an exploration, we organized a pilot mocap workshop with actors in spring 2025 at Humlab where we worked with two actors to try out a diverse set of prompts performed both individually and interactively. While some prompts were quick commands (Mandery et al. 2016), others were referring to affective moments such as attending a protest, sharing joy, walking while being surveilled, and so on. The session provided us with ethnographic insights from capturing mocap data with recognition of spatial, embodied, and technological dynamics. One emergent question was what is being lost in transferring bodies to data-bodies, such as how expressions and changes in emotions can or cannot be interpreted when we are only able to capture bodily movements. It is also important to consider the subjectivity and agency of actors in interpreting given prompts, and their interactions both with each other, the mocap system, facilitators, and the space. Mocap system also helps us to explore the affective role of bodies and bodily movements in storytelling practices, demonstrating alternatives to verbal storytelling, and to the usage of large language models (LLM).

References

Benjamin, R. (2019). Race after technology: abolitionist tools for the New Jim Code. Polity.

Braidotti, R. (2022). Posthuman feminism. Polity.

Buolamwini, J., & Gebru, T. (2018). Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. In Proceedings of the 1st Conference on Fairness, Accountability and Transparency, in Proceedings of Machine Learning Research (pp. 77-91)

Butler. J. (1990). Gender trouble: feminism and the subversion of identity. Routledge.

Càrdenas, M. (2016). Trans of color poetics: Stitching bodies, concepts, and algorithms. S&F Online, 13(3).

Chang, V. (2019). Catching the ghost: the digital gaze of motion capture. journal of visual culture, 18(3), 305-326.

Costanza-Chock, S. (2020). Design Justice: Community-Led Practices to Build the Worlds We Need (1st ed.). The MIT Press. https://doi.org/10.7551/mitpress/12255.001.0001

Danielsson, K., Tubella, A. A., Liliequist, E., & Cocq, C. (2023). Queer Eye on AI: binary systems versus fluid identities. In Handbook of Critical Studies of Artificial Intelligence (pp. 595-606). Edward Elgar Publishing.

Kilic, O. (2024). Lubunya Assemblages: Queer Networked Resistances in Turkey. Lund University Press.

Klipphahn-Karge, M., Koster, A. K., & dos Santos Bruss, S. M. (Eds.). (2023). Queer reflections on AI: uncertain intelligences. Taylor & Francis.

Liliequist, E., Tubella, A. A., Danielsson, K., & Cocq, C. (2023). Beyond the Binary - Queering AI for an Inclusive Future. interactions, 30(3), 31-33.

Mandery, C., Terlemez, Ö., Do, M., Vahrenkamp, N., & Asfour, T. (2016). Unifying representations and large-scale whole-body motion databases for studying human motion. IEEE Transactions on Robotics, 32(4), 796-809.

Muñoz, J. E. (2009). Cruising utopia: The then and there of queer futurity. New York University Press.

Nair, V., Guo, W., Mattern, J., Wang, R., O'Brien, J. F., Rosenberg, L., & Song, D. (2023). Unique Identification of 50,000+ Virtual Reality Users from Head & Hand Motion Data. arXiv preprint arXiv:2302.08927.

Puar, J. K. (2018). Terrorist assemblages: Homonationalism in queer times. Duke University Press.

Puar, J. K. (2020). “I would rather be a cyborg than a goddess”: Becoming-intersectional in assemblage theory. In Feminist theory reader (pp. 405-415). Routledge.

Russell, L. (2020). Glitch feminism: A manifesto. Verso Books.

Scheuerman, M. K., Paul, J. M., & Brubaker, J. R. (2019). How computers see gender: An evaluation of gender classification in commercial facial analysis services. Proceedings of the ACM on Human- Computer Interaction, 3(CSCW), 1-33.

Shew, A. (2023). Against technoableism: rethinking who needs improvement. WW Norton & Company.

Zytko, D., J. Wisniewski, P., Guha, S., PS Baumer, E., & Lee, M. K. (2022, April). Participatory Design of AI Systems: Opportunities and Challenges Across Diverse Users, Relationships, and Application Domains. In CHI Conference on Human Factors in Computing Systems Extended Abstracts (pp. 1- 4).

16:30–17:00 *LONG PAPER* **[67]**

Navigating Abundance: A Platform for AI-Augmented, Hybrid Reading of Multilingual Literary Texts

Sasha Rudan Kelbert[1], Eugenia Kelbert Rudan[2]

[1] *University of Oslo, Norway*

[2] *SAS, Slovakia*

Keywords: *Digital Humanities; Multilingual Corpora; AI-Augmented Reading; Retrieval-Augmented Generation (RAG); Knowledge Graphs; Literary Translation; Henrik Ibsen; Nabokov*

This paper presents the design and development of a research infrastructure for **AI-augmented hybrid reading** of multilingual literary corpora. The platform addresses the growing challenge of engaging critically and computationally with large, heterogeneous bodies of text that exist in multiple languages, versions, and translations. Our case studies include the **Henrik Ibsen multilingual corpus**—Norwegian originals and translations into English, German, Serbian, and other languages—and a **multilingual Nabokov corpus**, both hosted within the **LitTerra Digital Scholarly Environment (DSE)**. These corpora exemplify how abundance can obscure rather than reveal meaning: the wealth of digitized texts, translations, and metadata often outpaces our ability to interpret them. The platform responds to this problem by combining **Retrieval-Augmented Generation (RAG)**, **Large Language Models (LLMs)**, and **knowledge-graph-based metadata integration** to support research questions that move fluidly between humanistic interpretation and computational analysis.

1. Research Motivation

Digitization has created unprecedented access to world literature, yet multilingual and multi-version corpora remain difficult to explore coherently. A researcher investigating *A Doll House* might ask, *“How do different characters speak about money and finances across translations?”*—a question that requires not only textual retrieval but contextual understanding, sentiment interpretation, and cross-linguistic comparison. Current digital editions rarely enable such layered analysis. We therefore propose a **hybrid reading platform** that integrates human interpretive agency with LLM-driven assistance, enabling scholars to query, compare, and visualize literary phenomena across languages and historical layers.

2. Research Infrastructure and Architecture

The platform builds on an existing **graph-based knowledge infrastructure** (KnAllEdge) and the **LitTerra DSE**, extended with a suite of AI-enabled components:

Corpus Layer – Digitized and TEI-encoded texts, translations, and paratexts (letters, stage directions, criticism). Each text segment is annotated with language, speaker, and scene metadata.

Semantic Enrichment Layer – Automated **POS-tagging, named-entity recognition, sentiment analysis, and character-speech mapping** using open-source NLP models fine-tuned for Norwegian, English, and Slavic languages.

Vectorization and RAG Layer – Text segments are chunked and embedded using domain-specific embedding models. Vector representations are stored in **Weaviate** and cross-linked to the knowledge graph. Queries are processed through a **Retrieval-Augmented Generation pipeline**, allowing the LLM to reason over both semantic embeddings and structured metadata.

Hybrid Reading Interface – A multimodal reading environment that juxtaposes canonical text, translation, commentary, and AI-generated insight. Researchers can follow semantic references, explore automatically generated **graph visualizations** of themes and character networks, and trace how specific concepts evolve across translations.

3. Design Considerations and Methodological Challenges

(a) RAG Design and Vector Databases.

A core challenge is balancing linguistic fidelity with computational efficiency. We employ **multi-vector embeddings**—separate embeddings for surface text, semantic meaning, and contextual metadata—to preserve nuances of translation. **Weaviate's hybrid search** (vector + keyword) supports both precise citation retrieval and concept-based exploration. Custom tokenizers are used for morphologically rich languages (e.g., Norwegian Bokmål, Serbian).

(b) Integrating Additional Semantic Inputs.

Beyond text embeddings, we ingest **external analytical layers**: sentiment scores, rhetorical markers, stylistic indices, and intertextual references. These are represented as nodes and edges within the graph and exposed to the LLM through a **semantic-context API**. When generating responses, the LLM can cite or visualize relationships ("Helmer's speech in Act I shows stronger positive affect toward finances than Nora's in the English 1910 translation").

(c) Managing Output: From Text to Graph.

Instead of returning unstructured prose, the model outputs **structured semantic responses**—triples or JSON graphs—that reference corpus segments by URI. This allows visualization of argument flows, character relations, and sentiment trajectories. The result is not a replacement for close reading but an augmentation of it: the LLM acts as a *hermeneutic lens* embedded within the infrastructure.

4. Advanced Components and Extensibility

To orchestrate complex, multi-step queries, the system can employ **agent frameworks** such as **LangGraph** or the **Model Context Protocol (MCP)**.

LangGraph enables dynamic composition of reasoning chains: a user query ("trace economic metaphors in *A Doll House* across translations") triggers successive agents for retrieval, sentiment aggregation, and visualization.

MCP provides interoperability between our in-house services (NLP, metadata extraction, visualization) and external AI models, ensuring reproducibility and auditability of each step.

Even when all analytical services are hosted in-house, adopting these protocols supports **transparent orchestration, provenance tracking, and workflow reproducibility**—core requirements for sustainable digital-humanities infrastructures.

5. Research Directions and Use Cases

Comparative Character Studies.

Analyze how character agency and tone shift across languages. LLM-assisted queries reveal, for instance, variations in Nora's self-assertion across English and Serbian translations.

Diachronic Translation Analysis.

Trace lexical and syntactic evolution of key motifs (e.g., "freedom," "home," "money") across decades of translations.

Cultural and Ethical Framing.

Examine how translation choices reflect shifting moral and cultural perspectives—connecting sentiment analyses with historical metadata.

Readerly Interaction.

Enable interactive, guided exploration of themes via AI-generated narrative paths, supporting education and public humanities.

Cross-Corpus Generalization.

Apply the same pipeline to other multilingual corpora, such as Nabokov's self-translations, allowing comparative studies of authorial bilingualism and translingual aesthetics.

These scenarios serve not as isolated experiments but as **research proposals** demonstrating how AI-augmented reading can transform literary scholarship—from static edition to interactive semantic environment.

6. Discussion: From Abundance to Insight

The project directly addresses the DHNB 2026 theme *Lost in Abundance*. Massive digitization has produced plentiful but unevenly accessible literary data. Traditional digital editions privilege canonical texts, leaving peripheral translations and paratexts underexplored. Our approach treats this **abundance as an epistemic opportunity** rather than a burden. By embedding LLMs within transparent, citation-aware infrastructures, we can navigate large multilingual corpora while maintaining scholarly rigor and interpretability.

The platform thus contributes to current debates on:

Interpretability and bias in AI-assisted research;

Data provenance and citation in generative systems;

Methodological hybridity between close and distant reading; and

Infrastructure design for sustainable, open, and multilingual digital humanities.

7. Conclusion

The AI-augmented hybrid-reading platform represents a step toward **next-generation digital scholarly editions**—systems capable of dialoguing with texts rather than merely storing them. It demonstrates how RAG architectures, knowledge graphs, and agent frameworks can converge to support interpretive, multilingual, and participatory scholarship. Beyond Ibsen and Nabokov, the approach is extensible to any corpus characterized by multiplicity and translation, offering a methodological bridge between computational analysis and humanistic inquiry. Ultimately, it reimagines digital reading as an *interactive conversation with abundance*—one that respects linguistic diversity while leveraging the analytical power of AI.

17:00–17:30 *LONG PAPER* **[68]**

Computational Phonosemantics at Scale: Measuring Sound-to-Meaning Mappings in English and Polish with Gradient Boosting and GLMs

Szymon Pindur

Jagiellonian University in Krakow, Poland

Keywords: *phonosemantics, phonological iconicity, BERT fine-tuning, XGBoost, Generalized Linear Models*

Introduction

The present study undertakes a detailed empirical investigation of phonosemantic systematicity – statistically identifiable regularities in the mapping between phonological form and multidimensional experiential meaning – across the general vocabulary of English and Polish. Rooted in the classic

opposition between Saussurean arbitrariness (Saussure, 1966/1916) and Peircean iconicity (Peirce, 1931–1958), the idea of phonosemantic associations echoes the ancient *physei*/*thesei* debate on the natural versus conventional nature of the sign (Plato, 1961).

The core hypothesis addressed in the research posits that vocabulary in English and Polish manifests a certain degree of systematic associations between phonological units (phonemes, articulatory features, phoneme sequences) and a wide array of perceptual, sensorimotor, affective, and cognitive semantic dimensions, with partial cross-linguistic convergence driven by universal perceptual mechanisms. This aligns with a view that integrates iconicity and arbitrariness as complementary design features of language: the former facilitating semantic clustering, word learning, and cognitive efficiency, and the latter maximizing distinctiveness among semantically close items (Lockwood & Dingemanse, 2015; Dingemanse et al., 2015). Though direct sound imitation is largely absent from general vocabulary, analogical mappings – potentially rooted in articulatory gestures, acoustic properties of speech sounds (e.g., frequency, amplitude, periodicity), or perceptual entropy (Pindur, 2025) – may result in some systematic patterns.

Such mappings draw on experiential features shared across modalities: high/low pitch, abrupt/continuous onset, or tonal/aperiodic quality, which can in turn extend metaphorically to higher-order dimensions like valence, potency, or activity (Sidhu & Pexman, 2018). Several studies to date have addressed such phonosemantic correlations utilizing statistical methods. These include analyses of systematic sound-to-meaning mappings within single languages (e.g., Monaghan et al., 2014; Gutiérrez et al., 2016; or Winter & Perlman, 2021, who applied random forests to assess the simultaneous contribution of all English phonemes to perceived size), domain-specific classification tasks (e.g., Ngai et al., 2024, using XGBoost to predict gender from phonemes in Japanese names, revealing language-specific patterns such as /i/ → masculinity and /k/ → femininity), as well as cross-linguistic investigations of phonosemantic biases across thousands of languages (Blasi et al., 2016; Erben Johansson et al., 2020). Such work typically assumes that consistent cross-linguistic patterns in form-meaning associations reflect non-arbitrary, perceptually grounded regularities. While these approaches have revealed broad tendencies – particularly in size, shape, and emotional valence – they often rely on limited semantic categories and/or phonological features, leaving finer-grained experiential dimensions yet to be explored.

By constructing parallel phonosemantic profiles of English and Polish general lexis, our study aims to quantify the prevalence, strength, and cross-linguistic stability of these patterns, leveraging modern advances in NLP and data analysis. We thus strive to provide wider insight into the way speech sounds map onto a more diverse set of semantic dimensions in the general vocabulary of the two languages in question.

Materials and Methods

Our methodological framework combines distributional semantic modeling, high-dimensional phonological encoding, and predictive machine learning. Experiential semantics was operationalized using 65 continuous dimensions from Binder et al. (2016), spanning sensory, motor, spatial, temporal, affective, social, and cognitive domains. These dimensions are based on human ratings for 535 high-frequency English words and reflect neurobiologically grounded aspects of human experience.

To extend this framework to large-scale vocabularies, we fine-tuned transformer language models (BERT for English; HerBERT for Polish) with a regression head to predict the 65-dimensional semantic profiles. Trained and cross-validated on ~20,000 example sentences, the models achieved near-human performance (MSE ≈ 0.04; median R^2 ≈ 0.97), enabling semantic estimation for ~40,000 words per language.

Phonological representations were derived from standardized IPA transcriptions and encoded as sparse feature matrices capturing phoneme inventories, articulatory features, phoneme bigrams, morphological markers, and coarse POS categories.

Phonosemantic modeling employed XGBoost regression to identify the strongest predictors for each semantic dimension, followed by GLM estimation with Bonferroni correction. Effect sizes were quantified using Cliff's Delta. This pipeline allowed us to detect statistically robust sound–meaning correspondences across high-dimensional lexical data.

Results

The results reveal several consistent and statistically robust phonosemantic tendencies across both languages, particularly within sensory-perceptual and temporal domains.

a) English

In English, high front vowels (/ɪ/, /iː/) negatively predicted Vision-related dimensions, indicating reduced visual salience in words rich in these vowels. Nasals (/m/, /n/) positively predicted temporal and magnitude-related dimensions (e.g., Duration and Weight), while sibilants (/s/, /z/) positively predicted Audition and were associated with negatively valenced affect.

These distributed effects partially align with traditional phonesthemes. For example, the visual brightness cluster /gl-/ is compatible with the broader vowel–Vision structuring observed in the model, and /sn-/ clusters align with nasal involvement in action-related and affective domains. However, the results suggest probabilistic lexical tendencies rather than deterministic phonesthemic encoding: some canonical clusters (e.g., /fl-/ and fluid motion) did not emerge as primary phoneme-level predictors, indicating that such effects may rely on cluster-level or morphological structuring.

b) Polish

Polish bigram-level modeling revealed similarly structured phonosemantic effects, particularly in sensory-perceptual domains.

High front vowel and palatal/affricate clusters strongly negatively predicted Vision, Bright, and Color. In contrast, velar–low vowel sequences (e.g., /k/ + /a/) positively predicted visual dimensions. Bigrams containing nasal consonants positively predicted temporal dimensions, paralleling the English nasal–Duration tendency. Sibilant and affricate clusters positively predicted Audition, with Polish's richer sibilant inventory producing more differentiated contributions than in English.

Affective dimensions, however, displayed greater language-specific structuring. While sibilant involvement in negative affect was observable in English, Polish affective mappings were more diffusely distributed across palatal and affricate clusters, suggesting stronger modulation by phoneme inventory composition.

c) Cross-Linguistic Patterns

Cross-linguistic comparison reveals partial but systematic convergence. The strongest overlap appears in lower-level perceptual domains: high front vowels correlate with lower visual salience in both languages, nasal sequences contribute positively to temporal extension, while sibilant clusters correlate with higher auditory salience. In contrast, higher-order affective mappings show weaker cross-linguistic stability and greater sensitivity to language-specific phonological structure.

Taken together, the findings indicate that phonosemantic structure in general vocabulary is neither isolated nor purely language-specific. Instead, perceptually grounded tendencies appear to recur across languages, while their distribution is shaped by the structural affordances of individual phoneme inventories.

Contribution

The study aims to showcase large-scale computational phonosemantics as a part of digital humanities research utilizing modern data analysis methods. By scaling experiential semantic annotation through BERT-based regression, it eliminates reliance on costly human norming while achieving good generalization across wide vocabularies. By modeling ~40,000 words per language and integrating high-dimensional phonological representations, we show that phonosemantic structure remains detectable even within large and heterogeneous lexical inventories. In this sense, the study addresses linguistic abundance at three levels: lexical scale, semantic dimensionality, and phonological complexity. The results support a view of distributed iconicity: sound–meaning correspondences emerge as weak but systematic probabilistic biases embedded across large lexical systems. By situating these patterns within high-dimensional experiential space, the study extends phonosemantic analysis beyond traditional size, shape, and valence effects, demonstrating that subtle iconic structuring remains statistically detectable within abundant everyday vocabulary. Finally, by quantifying the phonosemantic patterns across neurobiologically grounded experiential dimensions, and comparing the findings with psycholinguistic insights, our study links the identified patterns with the human behavioral reality. In doing so, we hope to shed more light on the role that motivation and iconicity play in everyday language.

References

Anzarmou, Y., Mkhadri, A., & Oualkacha, K. 2022. “The Kendall interaction filter for variable interaction screening in high dimensional classification problems.” Journal of Applied Statistics, 50(7), 1496–1514.

Balota, D.A., Yap, M.J., Hutchison, K.A. et al. 2007. “The English Lexicon Project.” Behavior Research Methods 39, 445–459.

Batsuren, K., Bella, G., & Giunchiglia, F. 2021. “MorphyNet: a Large Multilingual Database of Derivational and Inflectional Morphology.” In Proceedings of the 18th SIGMORPHON Workshop on Computational Research in Phonetics, 39–48.

Binder, J. R., Conant, L. L., Humphries, C. J., Fernandino, L., Simons, S. B., Aguilar, M., & Desai, R. H. 2016. “Toward a Brain-Based Componential Semantic Representation.” Cognitive Neuropsychology, 33(3-4), 130–174.

Blasi, D. E., Wichmann, S., Hammarström, H., Stadler, P. F., & Christiansen, M. H. 2016. “Sound–meaning association biases evidenced across thousands of languages.” Proceedings of the National Academy of Sciences, 113(39), 10818–10823.

Carnegie Mellon Pronouncing Dictionary [CMUdict]. 2014. Carnegie Mellon University. Retrieved from http://www.speech.cs.cmu.edu/.

Chelba, C., et al. 2013. “One billion word benchmark for measuring progress in statistical language modeling.” arXiv preprint arXiv:1312.3005.

Devlin, J., Chang, M., Lee, K., & Toutanova, K. 2019. “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.” In Proceedings of the conference of the North American chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT), 4171–4186.

Dingemanse, M., Blasi, D. E., Lupyan, G., Christiansen, M. H., & Monaghan, P. 2015. “Arbitrariness, iconicity, and systematicity in language.” Trends in Cognitive Sciences, 19(10), 603–615.

Erben Johansson, N., Anikin, A., Carling, G., & Holmer, A. 2020. “The typology of sound symbolism: Defining macro-concepts via their semantic and phonetic features.” Linguistic Typology, 24(2), 253-310.

Gutiérrez, E. D., Levy, R., & Bergen, B. 2016. “Finding non-arbitrary form-meaning systematicity using stringmetric learning for kernel regression.” In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2379–2388).

Meissel, K., & Yao, E. S. 2024. “Using Cliff’s delta as a non-parametric effect size measure: an accessible web app and R tutorial.” Practical Assessment, Research, and Evaluation, 29(1).

Monaghan, P., Shillcock, R. C., Christiansen, M. H., & Kirby, S. 2014. “How arbitrary is language?.” Philosophical Transactions of the Royal Society B: Biological Sciences, 369(1651).

Mroczkowski, R., Rybak, P., Wróblewska, A., & Gawlik, I. 2021. “HerBERT: Efficiently Pretrained Transformer-based Language Model for Polish.” In Proceedings of the 8th Workshop on Balto-Slavic Natural Language Processing, 1-10.

Lockwood, G., & Dingemanse, M. 2015. “Iconicity in the lab: A review of behavioral, developmental, and neuroimaging research into sound-symbolism.” Frontiers in Psychology, 6, 1246.

Peirce, C.S. 1932. Collected Papers. Cambridge MA.

Pindur, S. 2025. “Sound symbolism in speakers of English: a qualitative synthesis.” International Journal of English Linguistics, 15(2).

Plato. 1961. “Cratylus.” In Hamilton E, Cairns H (Eds.) The collected dialogues. Princeton, NJ: Princeton University Press, 421-476.

Saussure, F. de. 1916/1966. Course in general linguistics. New York, NY: McGraw-Hill.

Sidhu, D. M., & Pexman, P. M. 2018. “Five mechanisms of sound symbolic association.” Psychonomic Bulletin & Review, 25, 1619–1643.

Turton, J., Vinson, D., & Smith, R. E. 2020. “Deriving contextualised semantic features from BERT (and other transformer model) embeddings.” In Proceedings of the 6th Workshop on Representation Learning for NLP (RepL4NLP-2021), 248–262.

Winter, B. & Perlman, M., 2021. “Size sound symbolism in the English lexicon.” Glossa: a journal of general linguistics 6(1): 79.

Wojdyga, G. 2018. “Results of the PolEval 2018 Shared Task 3: Language Models.” In Proceedings of the PolEval 2018 Workshop, 121-127.

Ylonen, T. 2022. “Wiktextract: Wiktionary as Machine-Readable Structured data.” In Proceedings of the 13th Conference on Language Resources and Evaluation (LREC), 1317-1325.

17:30–18:00 *LONG PAPER* **[69]**

Modeling Textual Emotions in Literary Fiction

Kirstine Nielsen Degn[1], Alexander Conroy[1], Xiaoyuan Jiang[2], Ali Al-Laith[1,2], Jens Bjerring-Hansen[1], Tanya Karoli Christensen[1], Ingo Zettler[3], Daniel Hershcovich[2]

[1] *Department of Nordic Studies and Linguistics, University of Copenhagen*

[2] *Department of Computer Science, University of Copenhagen*

[3] *Copenhagen Center for Social Data Science (SODAS), University of Copenhagen*

Keywords: *Computational literary studies, Emotion analysis, Sentiment analysis, Literary canons and archives, Literary historiography*

This paper explores how finetuned transformer models can be used to examine the vast and underexplored emotional-historical dimensions of literary history. Drawing on a corpus of 859 nineteenth-century Danish and Norwegian novels, we develop an interdisciplinary framework for analyzing the expression of emotions in literary texts. This paper presents three main contributions: (1) we introduce an annotated multilabel dataset of emotion categories for 859 nineteenth-century Danish and Norwegian novels; (2) we evaluate the performance and generalization capabilities of large language models in classifying emotions in long and complex texts; and (3) we propose a theoretical framework that integrates insights from psychology, linguistics, literary theory, and NLP to support future research on cultural and historical formations of emotions. By combining computational methods with explicit humanistic theorization, this approach not only enables large-scale analysis of literature's emotional dimensions but also demonstrates how methods used in the digital humanities can serve as a gateway to new theorization of emotions and their textual manifestations, offering valuable contributions to humanities scholars engaged in the so-called affective turn.

Session 4B — 16:00–17:30

16:00–16:30 **Lost in Structure: Graph-Based Technologies for Digital Scholarly Editions**
Sebastian Enns, Andreas Kuczera

16:30–17:00 **From Manuscript to Scripture: The Sanctification of N.F.S. Grundtvig's Writings**
Jon Tafdrup, Katrine Laigaard Baunvig

17:00–17:30 **What Happens to Style After Success? A Multidimensional Analysis of Context-Driven Expressive Drift in US-Published English-Language Novels**
Marc Barcelos

16:00–16:30 *LONG PAPER* **[70]**

Lost in Structure: Graph-Based Technologies for Digital Scholarly Editions

Sebastian Enns, Andreas Kuczera

TH Mittelhessen, University of Applied Sciences, Germany

Keywords: *Graph-based Technologies, Digital Editions, TEI, Scholarly editing, Interoperability*

Introduction

Digital scholarly editions have long relied on the Text Encoding Initiative (TEI) as their conceptual and technical backbone. While TEI has become a lingua franca for representing textual data, its hierarchical XML foundation imposes structural limits that increasingly clash with the complexity of contemporary editorial practice. The Ordered Hierarchy of Content Objects (OHCO) underlying TEI/XML cannot naturally represent overlapping hierarchies, discontinuities, or recursive annotation layers without resorting to workarounds that complicate encoding, processing, and querying (cf. Vogeler 2021; Stoyanova 2023). As a result, digital editions risk reducing the multiplicity of textual phenomena to the linear logic of markup, creating what Kuczera (2024) calls informational inaccuracy in interpretation.

This paper presents a graph-based alternative to hierarchical encoding through three complementary frameworks: Applied Text as Graph (ATAG), Explicitly Notated Citations (ENC), and Reusable Abstraction Model for Editorial Needs (RAMEN). Together, they outline an infrastructure for graph-based digital editions, which redefines text not as a tree of nested elements, but as a network of relations between textual, interpretive, and contextual entities. Within the theme of Lost in Abundance, these

models address the structural abundance of editorial data that is often lost within rigid markup hierarchies, proposing instead a graph-based ecology of editions that preserves complexity and interoperability.

From Hierarchies to Graphs

Building on earlier reflections within the TEI community, TEI Beyond XML (Kuczera 2022) develops ideas already anticipated by Cummings (2018), who noted that TEI's conceptual strength lies in its descriptive semantics rather than in its XML serialization. Detaching TEI concepts from the OHCO model allows textual data to be represented as networks of relations. This perspective assumes that research data in the humanities are fundamentally interconnected and that hierarchical organization should be applied only when it serves interpretation, not as a technical constraint.

However, XML still creates a gap between editorial reasoning and computational representation (cf. Pichler/Reiter 2022; Bode 2022). Complex textual features such as deletions, marginalia, or overlapping structures still require project-specific encodings that increase complexity and reduce interoperability. Previous work has shown that this proliferation of markup often obscures rather than clarifies editorial logic. Alternative approaches such as standoff markup (Schmidt 2016) and TagML (Bleeker et al. 2021) demonstrate that non-linear structures can be modeled without hierarchical containment, yet these systems remain technically fragile and rarely used in production. The central challenge, therefore, is to design frameworks that preserve TEI's descriptive richness while enabling flexible, editable, and semantically precise modelling beyond XML hierarchies.

Applied Text as Graph

Applied Text as Graph (ATAG) provides a structured solution to the limitations of hierarchical markup by modelling texts and annotations within a labeled property graph (Kuczera 2024). Each character is a uniquely identifiable node preserving reading order. Annotation nodes define their scope by linking to the first and last character of the relevant sequence, forming explicit relationships that allow overlaps and recursion. This architecture removes the need for nested markup and separates the linear text from its interpretive layers, addressing one of the central constraints of TEI/XML (cf. Enns et al. 2025).

In ATAG, editorial projects define the version of a text that serves as the main sequence, while alternative readings, comments, or corrections are attached as linked graph structures. Metadata can be integrated through TEI-conformant nodes, preserving descriptive richness while freeing TEI from the structural rigidity of XML (cf. Neill/Kuczera 2019; Schmidt 2016). This flexibility enables editing, versioning, and visualisation without breaking references.

The model has been tested in several digital editions, including the Liber Epistolarum (Dreyer et al. 2023) and Socinian Correspondence (Daugirdas/Kuczera 2017), where TEI/XML data were imported through a dedicated parser into an ATAG-based environment (Enns et al. 2025). In large-scale performance tests such as the import of over 200,000 Regesta Imperii (Kuczera et al. 2025) entries, graph databases like Neo4j handled more than 120 million nodes and edges without query loss, confirming the model's scalability. Through these implementations, ATAG establishes a bridge between traditional TEI encoding and graph-based infrastructures.

Explicitly Notated Citations

Explicitly Notated Citations (ENC) introduce a persistent and verifiable citation mechanism for digital editions that tackles the instability of current hyperlink and permalink systems (Enns/Kuczera 2025). Traditional web-based editions rely on HTML structures or positional indices to identify cited passages, which become unreliable when layout or content changes (cf. Koolen/Boot 2020). While permalinks and DOIs offer a partial solution, they generally refer to static resources and cannot represent dynamic text states or granular segments (cf. Bleier 2021; Stäcker 2020). Early approaches such as Canonical Text Services URNs (Tiepmar/Heyer 2017) provided persistent references for hierarchical corpora but remain unsuitable for networked or overlapping annotations. These limitations restrict the reuse of citations across platforms and undermine the long-term traceability.

ENC builds on ATAG by replacing positional references with unique character identifiers from the ATAG chain. Each citation records the UUIDs of the first and last characters in the referenced span, supplemented by a hash value encoding the entire sequence to ensure integrity verification. If the referenced text changes, the system can detect discrepancies and notify users. This structure allows citations to address not only the quoted passage but also its wider textual context, while remaining

independent of layout or document structure. By combining unique identifiers and hash validation, ENC establishes a stable, content-aware method of citation that aligns with the graph-based data model of ATAG and preserves the precise relation between text and annotation.

To ensure persistence and reproducibility, ENC integrates with common version-control workflows. The cited text and its annotations can be exported in machine-readable formats such as JSON and versioned through Git, with commit hashes appended to citation links (cf. Enns/Kuczera 2025). This enables earlier text states to be restored and referenced even after later editorial changes. In contrast to static citation systems, ENC supports both the publication and the ongoing creation of digital editions. It maintains reliable references even when texts remain in flux, providing stable connections throughout the editing process. By linking citation, provenance, and version management, ENC establishes an infrastructure in which dynamically evolving texts remain verifiable and addressable, facilitating editorial work as well as algorithmic analysis.

TEI in a Graph Ecosystem

The graph-based approach does not replace TEI but redefines its role within a broader data architecture. In an ATAG-based environment, TEI serves as a semantic vocabulary that structures descriptive information while relinquishing control over data organization. By abstracting TEI/XML into graph form, hierarchies can coexist, and interpretive statements become independent data entities (cf. Kuczera 2016, 2022, 2024). This allows for interoperable transformations between TEI and graph environments while maintaining consistency across editions (cf. Enns et al. 2025).

Such integration extends the concept of the assertive edition (cf. Vogeler 2019), in which editorial reasoning is explicitly modeled as data rather than embedded implicitly in markup. Graph imports of TEI further facilitate cross-edition linking, as annotations, entities, and citations can be shared as connected nodes within a broader scholarly network. This approach establishes a semantic layer that unites the descriptive vocabulary of TEI with the relational precision of graph models, aligning traditional encoding with the data practices of computational research.

RAMEN: A Multi-Level Metamodel for Reuse and Alignment

Reusable Abstraction Model for Editorial Needs (RAMEN) addresses the growing heterogeneity of digital editions, which now include texts, images, entities, annotations, and contextual relationships. Existing infrastructures such as TEI offer flexibility but often lead to project-specific models that are syntactically valid yet semantically incompatible (cf. Cummings 2019; Ciula et al. 2023). The result is a landscape of isolated, tailored models that hinder reuse and interoperability (cf. Liu et al. 2024). RAMEN resolves this issue through a shared abstraction layer that links project-specific logic to a common conceptual framework, ensuring transparency and comparability.

The framework introduces a multi-level architecture that unites modelling of and modelling for within a single system (McCarty 2018). It distinguishes between abstract metamodels and domain-specific layers, allowing inheritance and refinement across them. General, syntax-neutral concepts such as Text, Image, Entity, and Annotation can thus be instantiated according to the needs of individual projects without losing structural consistency. This model-driven approach is applied in projects like Socinian Correspondence (Daugirdas/Kuczera 2017), Liber Epistolarum (Dreyer et al. 2023), and Regesta Imperii Lab (Kuczera et al. 2025). By separating editorial logic from specific tools and formats, RAMEN maintains conceptual integrity across XML, JSON, and RDF environments, enabling interoperable and methodologically transparent digital editions.

Conclusion

Together, ATAG, ENC, and RAMEN outline a framework that moves digital scholarly editing from hierarchical markup toward relational and process-oriented modelling. In this perspective, editions are not static publications but evolving infrastructures that support both human interpretation and computational reuse. Representing texts, annotations, and citations as nodes within shared graphs enables provenance, referencing, and interoperability across projects.

Instead of reducing textual complexity to linear structures, graph-based models preserve and expose it as a meaningful network of relations. Treating TEI as a semantic vocabulary within this ecosystem allows existing standards to evolve into dynamic, algorithmically accessible environments. In doing so, graph-based editions transform abundance into structured knowledge and establish a sustainable foundation for future editorial research.

References

Bode, Katherine (2022): Doing (Computational) Literary Studies. New Literary History, vol. 54 no. 1, 2022, p. 531-558. Project MUSE. DOI: https://doi.org/10.1353/nlh.2022.a898320.

Bleeker, Elli / Dekker, Ronald Haentjens / Buitendijk, Bram (2021): Texts as hypergraphs: An intuitive representation of interpretations of text. Journal of the Text Encoding Initiative (14). DOI: https://doi.org/10.4000/jtei.3919.

Bleier, Roman (2021): How to cite this digital edition? DHQ: Digital Humanities Quarterly 15 (3). URL: https://digitalhumanities.org/dhq/vol/15/3/000561/000561.html.

Cummings, James (2018): A world of difference: Myths and misconceptions about the TEI. Digital Scholarship in the Humanities, 34 (Supplement_1), i58-i79. DOI: https://doi.org/10.1093/llc/fqy071.

Cummings, James (2019): Opening the book: data models and distractions in digital scholarly editing. International Journal of Digital Humanities 1 (2019): 179-193. DOI: https://doi.org/10.1007/s42803-019-00016-6.

Daugirdas, Kęstutis / Kuczera, Andreas (2017): Die sozinianischen Briefwechsel: Zwischen Theologie, frühmoderner Naturwissenschaft und politischer Korrespondenz. DFG-funded research project. (DFG project no. 324518514). URL: https://sozinianer.de.

Dreyer, Mechthild / Kuczera, Andreas / Stäcker, Thomas (2023): Das Buch der Briefe der Hildegard von Bingen. Genese – Struktur – Komposition. DFG-funded research project. (DFG project no. 530755431). URL: https://liberepistolarum.de.

Enns, Sebastian / Kuczera, Andreas (2025): Explicitly Notated Citations. Von digitalen zu algorithmischen Editionen. DHd 2025 Under Construction, Bielefeld, Deutschland. Zenodo. DOI: https://doi.org/10.5281/zenodo.15112014.

Enns, Sebastian / Armbruster, Stefan / Kuczera, Andreas (2025): Graph-Based Digital Editions and TEI, Towards Interoperable and Assertive Scholarly Editing. In: Book of Abstracts. New Territories. Text Encoding Initiative Conference and Members' Meeting 2025. September 16–20, 2025. Kraków, Poland. Zenodo. DOI: https://doi.org/10.5281/zenodo.17312233.

Koolen, Marijn / Boot, Peter (2020): Facilitating Reusable Third-Party Annotations in Digital Editions. Annotations in Scholarly Editions and Research: Functions, Differentiation, Systematization, edited by Julia Nantke and Frederik Schlupkothen, Berlin, Boston: De Gruyter, 2020, pp. 177-200. DOI: https://doi.org/10.1515/9783110689112-009.

Kuczera, Andreas (2022): TEI Beyond XML – Digital Scholarly Editions as Provenance Knowledge Graphs. In: Tara Andrews, Franziska Diehr, Thomas Efer, Andreas Kuczera and Joris van Zundert (eds.): Graph Technologies in the Humanities - Proceedings 2020, published at http://ceur-ws.org/Vol-3110.

Kuczera, Andreas (2024): Applied Text as Graph (ATAG). DHd 2024 Quo Vadis, Passau, Deutschland. Zenodo. DOI: https://doi.org/10.5281/zenodo.10698323.

Kuczera, Andreas / Pultar, Yannick / Kasper, Dominik / Prusova, Anna / Rübsamen, Dieter (2025): Regesta Imperii Online. Akademie der Wissenschaften und der Literatur Mainz. URL: https://www.regesta-imperii.de.

McCarty, Willard (2018): Modelling What There Is: Ontologising in a Multidimensional World. Historical Social Research/Historische Sozialforschung. Supplement 31: 33-45. DOI: https://doi.org/10.12759/hsr.suppl.31.2018.33-45.

Neill, Ian / Kuczera, Andreas (2019): The Codex – an Atlas of Relations. In: Die Modellierung des Zweifels – Schlüsselideen und -konzepte zur graphbasierten Modellierung von Unsicherheiten. Hg. von Andreas Kuczera / Thorsten Wübbena / Thomas Kollatz. Wolfenbüttel 2019. DOI: https://doi.org/10.17175/sb004_008.

Pichler, Axel / Reiter, Nils (2022): From Concepts to Texts and Back: Operationalization as a Core Activity of Digital Humanities. Journal of Cultural Analytics, vol. 7, no. 4, Dec. 2022, DOI: https://doi.org/10.22148/001c.57195.

Schmidt, Desmond Allan (2016): Using standoff properties for marking-up historical documents in the humanities, it - Information Technology, vol. 58, no. 2, pp. 63-69. DOI: https://doi.org/10.1515/itit-2015-0030.

Stäcker, Thomas (2020): A Digital Edition is Not Visible. Some Thoughts on the Nature and Persistence of Digital Editions. ZfdG - Zeitschrift für digitale Geisteswissenschaften, Nr. 5. Forschungsverbund Marbach Weimar Wolfenbüttel. DOI: https://doi.org/10.26083/tuprints-00019469.

Stoyanova, Silvia (2023): Articulating Intra- and Intertextual Relationships in the Fragment Collection. Working with the Digital Edition of Giacomo Leopardi's Zibaldone. magazén 4(1), S. 13-42. DOI: http://doi.org/10.30687/mag/2724-3923/2023/07/001.

Tiepmar, Jochen / Heyer, Gerhard (2017): An Overview of Canonical Text Services, Linguistics and Literature Studies, Vol. 5, No. 2, pp. 132 - 148, 2017. DOI: https://doi.org/10.13189/lls.2017.050209.

Vogeler, Georg (2021): Standing-off Trees and Graphs: On the Affordance of Technologies for the Assertive Edition. In: Graph Data-Models and Semantic Web Technologies in Scholarly Digital Editing,

herausgegeben von Elena Spadini, Francesca Tomasi und Georg Vogeler, S. 73-94. Norderstedt: Books on Demand (BoD). ISBN: 978-3-7543-4369-2.

16:30–17:00 *LONG PAPER* **[71]**

From Manuscript to Scripture: The Sanctification of N.F.S. Grundtvig's Writings

Jon Tafdrup, Katrine Laigaard Baunvig
Aarhus University, Denmark

Keywords: *scripturalisation, sacred text theory, canonization, Grundtvig, archival studies*

In this paper, we examine how the writings of N.F.S. Grundtvig (1783-1872), the Danish theologian, poet, and nation-builder, have undergone a remarkable transformation: from provisional, personal manuscripts to an authoritative corpus that is treated with reverence in both academic and popular contexts. Our central claim is that Grundtvig's works have been subject to a process we term *archival scripturalisation* – a dynamic in which secular literary and intellectual traditions acquire the structural and performative qualities typically associated with sacred texts.

Our inquiry is guided by three core questions:

1. Can the veneration of Grundtvig's archive be productively analysed through the lens of sacred text theory and scripturalisation studies?
2. How have institutional venues – especially the *Registrant over N.F.S. Grundtvigs papirer* (1957-1964), *Grundtvig-Studier* (1948-), and *Højskolebladet* (1876-) – shaped the canonization and exegetical framing of his writings?
3. What insights can computational methods offer into the discursive sanctification of Grundtvig across different historical periods and genres of reception?

We approach the digitized corpora of *Registrant over N.F.S. Grundtvigs papirer* (1957-1964), the two journals *Grundtvig-Studier* (1948-), and *Højskolebladet* (1876-) as interconnected infrastructures that both preserve and conceal materials within abundance. By applying computational approaches such as keyword modelling, embeddings, and semantic network analysis, we trace recurring vocabularies of reverence, authority, and sanctity across these sources, contributing to emerging computational approaches to modelling canon formation and literary prestige in reception history. This allows us to extract and map what we call "terms of scripturalisation," showing how editorial routines and citation loops confer cultural scripture status on Grundtvig. In doing so, we identify how canonization emerges not only from what is highlighted, but also from what is systematically backgrounded or institutionally sidelined in order to maintain the coherence of a culturally sanctified authorship within these digitized collections (Bode 2017).

Our research grows out of an interdisciplinary dialogue between Religious Studies, Archival Studies, and the Digital Humanities. At stake is not only the reception of Grundtvig himself, but also the broader question of how national figures are canonized through textual and archival practices. We argue that Grundtvig's afterlife exemplifies how archives and institutions can confer sanctity upon texts, positioning them as cultural scripture (Baunvig 2023). The case of Grundtvig offers a striking example of how modern archives can take on ritual, theological, and even relic-like functions in a supposedly secular cultural field. Accordingly, we hypothesize that Grundtvig's textual corpus has indeed undergone sacralization, both structurally and performatively, within Danish cultural and intellectual history.

Theoretical Framework: Sacredness and the Archive

To conceptualize the author's archive in a sacrosanct context, we must go beyond simply narrating the history of archives and their emergence. Instead, we situate our work within what has been termed the *archival turn* – a profound shift in thinking about the form and function of the archive inaugurated by Michel Foucault and later developed by Jacques Derrida (Foucault 2002; Derrida 1996) and subsequently taken up across the social sciences and humanities. The purpose of this theoretical turn is to develop an operational vocabulary capable of capturing the author's archive in its multifaceted essence – as something more than a mere collection of documents (in this case, manuscripts) physically attached to their canonical originator.

The modern idea of a writer's archive begins, in many ways, with a transformation in authorial self-consciousness: a growing impulse to preserve the burgeoning production of texts and to assemble a testimony for posterity. In Grundtvig's case, the earliest manuscripts date from 1798, when he was only fifteen years old. Yet Grundtvig, of course, was no isolated figure: he was part of a broader European movement already several decades in the making, where the preservation of manuscripts became bound up with questions of authorship, originality, and cultural heritage. Christian Benne has described this shift in mentality in six distinct phases, of which the first three help illuminate how the archive may be understood as a canonical and ritual artifact (Benne 2021).

In this investigation, we bracket the question of what purpose Grundtvig himself may have had in mind for his papers. Instead, we begin with the simple fact of their existence: at the time of his death, Grundtvig left behind a manuscript archive of approximately 45,000 folios. It is this massive, physical corpus that later institutions and interpreters transformed into an object of veneration – treated not merely as evidence of a writer's life, but as a relic imbued with cultural sanctity.

This perspective allows us to combine sacred text theory and archival theory. From the anthropology of religion, we use Roy Rappaport's claim that texts become ritually stabilized through repetition and institutionalisation (Rappaport 1999). From Vincent Wimbush, we adopt the concept of *scripturalisation*, which highlights how corpora achieve power not only through content but through their ritualised use and cultural positioning (Wimbush 2012). Derrida's *Archive Fever* alerts us to the hermeneutic privilege of the archivist, while Foucault's reflections on discourse and classification clarify how cataloguing practices generate new epistemic regimes. Together, these theories frame Grundtvig's archive as both a religious and an archival phenomenon.

Data

We focus on three major sites of Grundtvig reception, each of which demonstrates a distinct mode of textual sacralization.

- **The *Registrant over N.F.S. Grundtvigs papirer*** (1957-1964), a thirty-volume inventory of Grundtvig's manuscripts, effectively canonized his writings. By classifying genres and ordering the corpus, the *Registrant* imposed interpretative limits and created a textual body analogous to a religious canon. Just as the biblical canon was stabilized by ecclesiastical councils, Grundtvig's writings were stabilized by editorial decisions that determined what counted as his "corpus" and what could be set aside.
- **The journal *Grundtvig-Studier*** (1948-) functions like an exegetical tradition. Its cycles of commentary, citation, and quotation parallel rabbinic, patristic, or scholastic traditions that maintain orthodoxy through ongoing interpretation. This repetitive engagement not only consolidates Grundtvig's intellectual authority but also produces a sense of ritualized study, where scholars participate in a form of Grundtvigian devotion, even if secularized.
- **The journal *Højskolebladet*** (1876-) represents the popular dimension of canonization. Grundtvig's words are mobilized in national and cultural discourse, portraying him as a prophetic figure whose writings ground collective identity. This dynamic bears comparison to other national "scriptures," such as Martin Luther in German Protestant nationalism or Thomas Jefferson in American civil religion. *Højskolebladet* illustrates how textual sanctity is not limited to elite scholarly circles but is diffused through public, cultural, and educational discourse.

Taken together, these three institutions demonstrate how Grundtvig's works circulate across registers – archival, scholarly, and popular – each reinforcing their sanctity through different discursive strategies.

Analysis: Institutionalizing Textual Authority

While our argument is grounded in theory and archival case studies, we employ computational methods to trace long-term shifts in Grundtvig's textual sanctification. Using periodised word embeddings and semantic network analysis, we analyse discursive formations across multiple historical time slices in corpora drawn from *Grundtvig-Studier* and *Højskolebladet*. These techniques allow us to identify the changing semantic neighbourhoods around "Grundtvig" and related terms, making visible the shifting contours of his reception across decades (Bode 2017; Baunvig 2023).

Our preliminary analyses suggest:

- A gradual intensification of ritualistic and reverential language in *Grundtvig-Studier*, particularly after 1948, when scholarship on Grundtvig took on a more hagiographic tone.

- Divergent semantic clusters in *Højskolebladet*, where Grundtvig is consistently associated with national, prophetic, and pedagogical themes, distinct from the academic framing.
- Temporal variation in discourse, reflecting broader cultural shifts – from the romantic nationalism of the late 19th century to the more secularized yet still reverential readings of the late 20th century.

These empirical insights complement our theoretical framing, offering a novel way to measure and visualize the sacralization of a secular corpus.

Conclusion and Contribution

We conclude by proposing the concept of *archival scripturalisation*: the process by which secular archives come to function structurally and performatively like sacred canons. Grundtvig's case demonstrates how textual veneration operates across institutional, scholarly, and popular domains, blurring the boundary between secular and sacred.

Our contribution is twofold: first, to Religious Studies, by extending sacred text theory beyond explicitly religious traditions; and second, to Digital Humanities, by demonstrating how computational methods can illuminate processes of canonization and cultural sanctification that unfold across long temporal spans.

References

Benne, Christian. 2021. "'Ingen indfald bør gå tabt': Det moderne arkiv og den litterære selvrefleksion." In Danske forfatterarkiver, edited by Anders Juhl Rasmussen and Thomas Hvid Kromann, 35–55. Odense: Syddansk Universitetsforlag.

Baunvig, Katrine Frøkjær. 2023. "'Each of Our Springs Has Lost Its Miraculous Power': The Range of a Religious Hotspot – A Distant Reading of Lourdes Representations in Denmark, 1858–1914." Numen 70 (1): 43–69.

Bode, Katherine. 2017. "The Equivalence of 'Close' and 'Distant' Reading; or, Toward a New Object for Data-Rich Literary History." Modern Language Quarterly 78 (1): 77–106.

Derrida, Jacques. 1996. Archive Fever: A Freudian Impression. Translated by Eric Prenowitz. Chicago: University of Chicago Press.

Foucault, Michel. 2002. The Archaeology of Knowledge. Translated by A. M. Sheridan Smith. London: Routledge.

Rappaport, Roy A. 1999. Ritual and Religion in the Making of Humanity. Cambridge: Cambridge University Press.

Wimbush, Vincent L. 2012. White Men's Magic: Scripturalization as Slavery. New York: Oxford University Press

17:00–17:30 *LONG PAPER* **[72]**

What Happens to Style After Success? A Multidimensional Analysis of Context-Driven Expressive Drift in US-Published English-Language Novels

Marc Barcelos

TEXT: Center for Contemporary Cultures of Text, Aarhus University, Denmark

Keywords: *Novels; Computational Stylistics; Multidimensional Stylistic Analysis; Semantic Embeddings*

The field of stylistics was formalized in 1909, motivated by a larger systemic trend away from the intuition-based exploration of style in text to a more formal analysis of linguistic patterns and their impact on a writer's stylistic identity[1]. Across disciplines, the perception of writer identity has since shifted over time from an inherent, aesthetic property of one's works towards a dynamic, contextually informed, and cognitively driven understanding of textual style[2]. Early approaches relied upon the claim that a text's meaning and style was intrinsic, whereas contemporary approaches implicate personal variation, cognition, social context, and empirical analysis to define writer identity more holistically[3]. Together, these developments represent a broad interdisciplinary move towards flexible, process-based approaches to how writer identity is produced, developed, and interpreted[4].

This project builds upon this current understanding of writer identity, suggesting that writer identity is inherently dynamic and emergent from one's aims, previous experience, current internal state, and changing external contexts. As intuition suggests, writers tend to work within stylistic proximity to their previous work[5]. However, this similarity may not be a result of an inherent feature of an individual or their work, but instead the collaboration of their subjective state with the task set before them[6]. Thus,

this project frames each text as a measurable expression of one's shifting constellation of internal variables and contextual constraints. These expressions aggregate into what is introduced in this project as an expressive profile, the emergent patterns that shape and define how writers mobilize language across contexts.

To translate this conceptual approach, the study utilizes a corpus-based design that treats empirical measurements of style in novels as snapshots of an author's expressive profile at different moments in their career, allowing for direct comparison of how expression shifts when an author's cultural position changes[7]. The analysis, therefore, will focus on variation across career phases defined by sudden changes in reception, namely after their first commercial breakthrough (appearing on a major bestseller list) or institutional recognition (major literary award recognition), treating these moments as contextual inflection points. The corpus of novels to be explored consists of English-language novels published in the United States from 1880-2000[8]. For each author, works are sampled immediately before and after recognition periods to provide a balanced view of writing before and after shifts in visibility and prestige. This structure enables a controlled investigation of how changes in audience scale, market positioning, and institutional valuation influence expressive behavior.

Because these authors continue to write novels even as their reception contexts evolve, the study can examine stylistic adaptation without the confounding effects of switching form or domain. This relative stability, paired with measurable changes in readership and recognition, creates a natural setting for observing how external pressures reshape individual expression. Authors may address similar thematic concerns across their careers, yet their expressive profiles can diverge as new expectations recalibrate their expressive profile. From this perspective, expressive profiles are treated as dynamic configurations that respond to changing communicative environments. Style emerges as situated and continually renegotiated as writers move through different contexts, revealing the dynamics of how literary expression is shaped by both individual disposition and the evolving contexts in which it is produced and received.

Operationalizing this dynamic nature of expression requires an analytical framework capable of capturing patterned variation across many linguistic features simultaneously. Multidimensional analysis (MDA), pioneered by Biber (1988), does precisely this. MDA uses principal component analysis (PCA) to model style not as a single feature or marker, but as multiple clusters of features whose patterned correlations reveal functional dimensions of expression[9]. Within this framework, expression is understood as a constellation of co-occuring linguistic choices that reflect specific communicative purposes and situational contexts[10]. MDA approaches have been widely applied to distinguish registers, genres, and discourse types, demonstrating how functional variation emerges from systematic linguistic patterning rather than isolated stylistic traits [11]. However, prior MDA research has primarily been utilized in focused applications rather than at scale across large corpora. The present study extends MDA by analyzing expressive profiles as dynamic, hierarchical phenomena spanning multiple analytical levels, from individual texts to author-level patterns and group-level trends. Contextual inflections are explicitly modeled by dividing each author's body of work into an equal number of texts pre- and post-recognition and noting the number of publications removed from the inflection point of a particular work.

At the same time, the project expands MDA's traditional feature set by integrating contemporary semantic modeling techniques alongside established linguistic measures. Using TextDescriptives[12] and E5 text embeddings[13], each text is quantified through 121 linguistic and semantic metrics capturing syntactic signals (such as sentence complexity, dependency patterns, lexical diversity), discourse-level properties (such as cohesion, information flow, readability), and embedding-based measures that model semantic similarity and coherence across sentences, passages, and authors. The combination allows style to be captured not just as word choice or grammar, but as the interaction between form, processing difficulty, and conceptual progression through a text. Through this, MDA becomes not only a method for identifying expressive variation but a scalable framework for modeling one's expressive profile as a contextualized, multi-layered pattern. Expressive profile, in this sense, operationalizes the relationship between form and function while providing a computational means of tracing how individual expressive habits interact within evolving environments.

The study expects that following moments of major recognition, either from the public via the market or the field via major awards, one's expressive profile in future works will display a divergent pattern from prior works in respond to their altered standing. When a work achieves bestseller status or receives institutional recognition, an author's context of production can change abruptly. Audience scale expands, expectations become more explicit, and the author's position within the literary field is

adjusted. These inflection points are therefore hypothesized to impact writers in ways that lead them to reorganize their expressive behavior. Some authors may shape all future work to remain stylistically close to the recognized work, consolidating features associated with their newfound recognition, while others may react in the opposite direction, producing texts that are measurably distinct as they try to redefine their expressive profile or resist emerging expectations.

Empirically, an increase in stylistic dispersion rather than convergence across post-recognition works is expected across authors. Pre-inflection texts are likely to show tighter clustering, reflecting relatively stable conditions, whereas post-recognition texts may occupy a broader range of metrics and expressive patterns. Rather than exhibiting a uniform shift, authors are anticipated to respond in varied ways, with some consolidating their most present previous stylistic tendencies, while others experimentally oscillate between or gradually shift toward new expressive patterns. This would, taken together, suggest that writer identity is not a fixed signature carried consistently across time, but an emergent configuration shaped by the interaction of prior habits, shifting audiences, and evolving cultural pressures. Furthermore, this would support a dynamic model of expressive profiles in which style operates as an adaptive system responding to contextual change rather than as an intrinsic, unchanging property of authorship. Recognition, in this view, becomes a catalyst for expressive repositioning, revealing how literary identity is continually renegotiated as writers move through different social and institutional environments.

This study will help in continuing to formalize quantitative stylistic measurement as a method for studying literary expression as an evolving practice rather than a stable signature. By modeling expressive profiles across moments of shifting reception, the study demonstrates how linguistic and semantic metrics can explore the interactions between authorial agency and historically situated constraints such as audience expansion, institutional validation, and market visibility. Each metric captures a different dimension of expressive choice, allowing researchers to assemble temporally grounded snapshots of how writers recalibrate voice, structure, and semantic orientation across their careers. Comparing these snapshots across professional and cultural contexts makes it possible to trace not only stylistic change, but how it unfolds within the literary field. In this way, the project extends multidimensional analysis beyond classification or attribution, using it instead to examine writing in literary practices as an adaptive, career-long process embedded within internal and external systems of reception and valuation.

This process-oriented operationalization of writer identity through the use of expressive profiles also suggests new methodological pathways for computational stylistics more broadly. If expressive profiles can be modeled as dynamic configurations shaped by context, then the same metrics can be used to explore how contemporary writing technologies, namely generative artificial intelligence (GAI), participate and modulate similar processes of negotiation and constraint. Rather than treating computational tools as external to authorship and literary production, this framework includes them within the conceptualization that has long linked authors, audiences, and institutions. From this perspective, GAI can be studied as more than simply a producer of text, but as a different context in which expressive parameters are adjusted, tested, and readjusted. Such an approach provides a new way to examine human and machine-assisted writing under the same analytical lens, reframing authorship in the age of AI as an extension of the historically ongoing relationship between creative intention and cultural context.

References

[1] (Berenike Herrmann et al., 2015; Giovanelli & Harrison, 2022; McIntyre & Busse, 2010)

[2] (Genova, 1979; Hirsch Jr, 1975; Hyland, 2010; Juola et al., 2019; Peng & Hengartner, 2002; Wilson et al., 2022)

[3] (Aristoteles et al., 2009; Kayser, 1948; Meyer, 1956)

[4] (Carter et al., 2008; Stockwell, 2020; Toolan, 2019)

[5] (Amancio, 2015; Johnson & Wright, 2014; Seminck et al., 2022)

[6] (Hughes et al., 2012; Lin et al., 2016; Schwartz et al., 2017; Wang & Hu, 2018)

[7] (Conrad, 2002)

[8] (Bizzoni et al., 2024)

[9] (Bro & Smilde, 2014; Nokeri, 2021)

[10] (Biber & Conrad, 2019)

[11] (Biber, 1995; Conrad, 2002)

[12] (Hansen et al., 2023)

[13] (Wang et al., 2024)

Amancio, D. R. (2015). Authorship recognition via fluctuation analysis of network topology and word intermittency. Journal of statistical mechanics, 2015(3), P03005–03020. https://doi.org/10.1088/1742-5468/2015/03/P03005

Aristoteles, Ross, W. D., & Brown, L. (2009). The Nicomachean ethics. Oxford University Press.

Berenike Herrmann, J., van Dalen-Oskam, K., & Schöch, C. (2015). Revisiting Style, a Key Concept in Literary Studies. Journal of literary theory (Berlin), 9(1), 25–52. https://doi.org/10.1515/jlt-2015-0003

Biber, D. (1988). Variation across Speech and Writing. Cambridge University Press. https://doi.org/DOI: 10.1017/CBO9780511621024

Biber, D. (1995). Dimensions of Register Variation: A Cross-Linguistic Comparison. Cambridge University Press. https://doi.org/DOI: 10.1017/CBO9780511519871

Biber, D., & Conrad, S. (2019). Register, Genre, and Style (2 ed.). Cambridge University Press. https://doi.org/DOI: 10.1017/9781108686136

Bizzoni, Y., Feldkamp, P., Lassen, I. M. S., Thomsen, M. R., & Nielbo, K. L. (2024, May). A Matter of Perspective: Building a Multi-Perspective Annotated Dataset for the Study of Literary Quality. In N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti, & N. Xue, Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024) Torino, Italia.

Bro, R., & Smilde, A. K. (2014). Principal component analysis [10.1039/C3AY41907J]. Analytical Methods, 6(9), 2812–2831. https://doi.org/10.1039/C3AY41907J

Carter, R., Stockwell, P., Stockwell, P., & Carter, R. (2008). Stylistics: retrospect and prospect. In (1 ed., pp. 291–302). Routledge. https://doi.org/10.4324/9781003060789-32

Conrad, S. (2002). Corpus Linguistic Approaches for Discourse Analysis. Annual Review of Applied Linguistics, 22, 75 – 95.

Genova, J. (1979). The Significance of Style. The Journal of Aesthetics and Art Criticism, 37(3), 315–324. https://doi.org/10.2307/430785

Giovanelli, M., & Harrison, C. (2022). Stylistics and Contemporary Fiction. English Studies, 103(3), 381–385. https://doi.org/10.1080/0013838X.2022.2043035

Hansen, L., Olsen, L. R., & Enevoldsen, K. (2023). TextDescriptives: A Python package for calculating a large variety of metrics from text. arXiv.org. https://doi.org/10.48550/arxiv.2301.02057

Hirsch Jr, E. (1975). Stylistics and synonymity. Critical Inquiry, 1(3), 559–579.

Hughes, J. M., Foti, N. J., Krakauer, D. C., & Rockmore, D. N. (2012). Quantitative patterns of stylistic influence in the evolution of literature. Proceedings of the National Academy of Sciences - PNAS, 109(20), 7682–7686. https://doi.org/10.1073/pnas.1115407109

Hyland, K. (2010). Community and Individuality: Performing Identity in Applied Linguistics. Written communication, 27(2), 159–188. https://doi.org/10.1177/0741088309357846

Johnson, A., & Wright, D. (2014). Identifying idiolect in forensic authorship attribution : an n-gram textbite approach. Language and Law, 1(1).

Juola, P., Mikros, G. K., & Vinsick, S. (2019). Correlations and Potential Cross-Linguistic Indicators of Writing Style. Journal of quantitative linguistics, 26(2), 146–171. https://doi.org/10.1080/09296174.2018.1458395

Kayser, W. (1948). Das sprachliche Kunstwerk : eine Einführung in die Literaturwissenschaft. Francke.

Lin, Y. R., Margolin, D., & Lazer, D. (2016). Uncovering social semantics from textual traces: A theory-driven approach and evidence from public statements of U . S . M embers of C ongress. Journal of the Association for Information Science and Technology, 67(9), 2072–2089. https://doi.org/10.1002/asi.23540

McIntyre, D., & Busse, B. (2010). Language and Style (1 ed.). Bloomsbury Publishing Plc.

Meyer, H. (1956). Die Kunst der Interpretation. Studien zur deutschen Literaturgeschichte. In (Vol. 30, pp. 418–418): University of Oklahoma Press.

Nokeri, T. C. (2021). Principal Component Analysis with Scikit-Learn, PySpark, and H2O. Data Science Solutions with Python.

Peng, R. D., & Hengartner, N. W. (2002). Quantitative Analysis of Literary Styles. The American statistician, 56(3), 175–185. https://doi.org/10.1198/000313002100

Schwartz, R., Sap, M., Konstas, I., Zilles, L., Choi, Y., & Smith, N. A. (2017). The Effect of Different Writing Tasks on Linguistic Style: A Case Study of the ROC Story Cloze Task. https://doi.org/10.48550/arxiv.1702.01841

Seminck, O., Gambette, P., Legallois, D., & Poibeau, T. (2022). The Evolution of the Idiolect over the Lifetime: A Quantitative and Qualitative Study of French 19th Century Literature. Journal of Cultural Analytics, 7(3). https://doi.org/10.22148/001c.37588

Stockwell, P. (2020). Language and Literature. The Handbook of English Linguistics.

Toolan, M. (2019). Literary Stylistics. In: Oxford University Press.

Wang, J., & Hu, C. (2018). Similarity, Metaphor and Creativity. Language and Semiotic Studies, 4(3), 101–116. https://doi.org/doi:10.1515/lass-2018-040306

Wang, L., Yang, N., Huang, X., Yang, L., Majumder, R., & Wei, F. (2024). Multilingual E5 Text Embeddings: A Technical Report. https://doi.org/10.48550/arxiv.2402.05672

Wilson, R., Bhandarkar, A., & Woodard, D. (2022). TraSE: Towards Tackling Authorial Style from a Cognitive Science Perspective. https://doi.org/10.48550/arxiv.2206.10706

Session 4C — 16:00–17:40

16:00–16:30 **”The Multimodal Television Generation”: Experimental Synchronization in Early U.S. and Swedish Children’s Television**
Johan Malmstedt

16:30–17:00 **Tracing Terrorism in Television: Toward a Methodology of Large-Scale Audiovisual Search**
Daniel Brodén, Johan Malmstedt, Mats Fridlund

17:00–17:20 **Scene-Anchored Analysis of Affective “Stickiness” in Nordic Political-Thriller Television with Large Language Models**
Aida Gholami

17:20–17:40 **Framing Digital Transformation: Media Discourses on Digitalization in Sweden**
Coppélie Cocq, Stefan Gelfgren, Rebecka Weegar

16:00–16:30 *LONG PAPER* **[73]**

”The Multimodal Television Generation”: Experimental Synchronization in Early U.S. and Swedish Children’s Television

Johan Malmstedt
University of Gothenburg, Sweden

Keywords: *multimodal communication, children’s television, cross-modal literacy, media history, sound studies*

Children’s television of the 1970s remains a debated and often controversial field, sometimes celebrated as a site of creative experimentation and progressive pedagogy, and at other times dismissed as overly didactic and moralizing (Mares and Pan, 2013; Janson, 2014). This decade nevertheless marked a turning point in how audiovisual media addressed young audiences across both Europe and the United States, merging education and entertainment in ways that continue to shape contemporary media culture. While cases like *Sesame Street* (PBS, 1969) and *Fem myror är fler än fyra elefanter* (SVT, 1973–1975) have been celebrated for their measurable learning outcomes and aesthetic innovation (Mares and Pan, 2013; Djerf-Pierre and Weibull, 2001), most scholarship has focused on policy, institutional history, and measurable effects, leaving the audiovisual mechanisms of learning comparatively underexplored.

In the early 1970s, communication scholar Ray Birdwhistell warned that the emerging “television generation” was learning that communication is a multimodal process, one that makes use of a multitude of auditory and visual cues in interpreting a message. He also argued that producers carried a special responsibility to maintain the relationship between sound and image that television had come to model. Taking Birdwhistell’s insight as a point of departure, this study asks how children’s television of the 1970s arranged the relationships between sound and sight. It examines how these programs taught young viewers to interpret meaning across modes, and how verbal and visual codes were combined to stabilize meaning for novice audiences. By using a comparative analysis between Sweden and the United States, it becomes possible to explore how contrasting public service and commercial conditions shaped these multimodal pedagogies. Building on semiotic studies of children’s television (Hodge and Tripp, 1986; Fiske, 1987) and theories of audiovisuality (Altman, 1985; Chion, 1994; Donnelly, 2019), I approach these series as early laboratories for what I call cross modal literacy.

Children’s Television and the Shaping of Aesthetic Experience

By the early 1970s, both nations had committed to new forms of children's programming that combined education, creativity, and entertainment. As Norma Pecora observed, producers became acutely concerned with "the ways in which children of different ages make sense of and utilize the messages of television" (Pecora, 2002: 3).

In the United States, the founding of PBS and the Carnegie Commission's call for "educational excellence" led to shows like *Sesame Street* and *Zoom*, which integrated developmental psychology with popular culture (Mares & Pan, 2013). In Sweden, Sveriges Television (SVT) pursued similar ambitions under a public-service ethos, developing programs such as *Anita och Televinken* and *Fem myror är fler än fyra elefanter*, which emphasized linguistic play, equality, and social participation (Janson, 2014; Scannell, 1996).

While American programming balanced public-service ideals against commercial competition, Swedish television maintained a distinctly non-commercial orientation. Yet both systems faced the same aesthetic challenge: how to bridge abstraction and concreteness in children's comprehension of audiovisual narratives. As Hodge and Tripp (1986) noted, young viewers depend on visuals to "concretize" meaning, while verbal language orders time and logic. Effective pedagogy, therefore, required the integration of sound and image, the kind of "audiovisual contract" later theorized by Chion (1994), to sustain clarity and engagement.

Source Material and Crossmodal Analysis

To investigate these dynamics, I compiled a comparative corpus of 1970s children's television from Sweden and the United States, drawing on digitized archives from the Svensk Mediedatabas and the American Archive of Public Broadcasting. The sample includes *Sesame Street*, *The Electric Company*, and *Zoom* (U.S.), alongside *Anita och Televinken*, *Fem myror är fler än fyra elefanter*, and *Vilse i Pannkakan* (Sweden).

To capture what is shown and how it is sounded and edited, I combine transformer-based recognition with calibrated measures of audiovisual complexity. For the visual stream, I use Moondream, a vision transformer with high accuracy object detection capacity. Each frame is analyzed for object presence and recurrence and ultimately mapped against objects in the audio stream. For the soundtrack, I employ an Audio Spectrogram Transformer (AST) to identify events like speech, song, counting, clapping, and musical stingers. Aligning these cues with visual detections allows me to study cross-modal congruence: when a spoken label coincides with its visual referent, or lead–lag relations where sound anticipates or follows image changes.

I also compute calibrated 0–100 complexity curves for both sound and image using fixed, domain-anchored mappings, ensuring comparability across titles (Fickers et al., 2019; Malmstedt, 2023). The visual analysis captures spatial–temporal and motion features such as shot changes, optical flow, and edge density, while the audio analysis models loudness variation, spectral entropy, and rhythm. These signals identify moments of heightened multimodal intensity and patterns of synchrony that inform close readings of editing, rhythm, and pedagogical pacing, allowing me to trace how these methods vary across episodes and across national contexts

Findings and Discussion

Preliminary findings reveal two recurring and comparative strategies: cross-modal redundancy, where objects content and information complexity work in synchrony; sensory cueing, where one modality briefly diverges to prepare anticipation for a more comprehensive change in content. Furthermore, the results show that cueing mechanisms have a balanced distribution between the visual and audio domain. This challenges the assumption that audiovisual media tend towards being visually driven. Across both Swedish and U.S. programs, it is possible find frequent cases where sound leads the visual field, melodic shifts, vocal stress, or rhythmic cues anticipating cuts or object reveals, indicating that sonic structure often organizes perception rather than simply reinforcing what is seen.

This finding suggests more openness to modal non-synchronicity than usually ascribed to this period of audiovisual culture. Previous scholarship has argued that extensive experiments in argued that experimentation Techniques related to intentionally not aligning sound and visual elements have been attributed to the later integration of digital work environments and tools (Donnelly, 2019). However, my results demonstrate how children's television of the 1970s was already experimenting with unorthodox sound–image relationships. The pedagogical coordination of rhythm, speech, and editing in these

programs demonstrates that cross-modal design emerged long before contemporary digital postproduction or algorithmic sound synchronization.

Conclusions

This article reframes 1970s children's television as an experimental arena for audiovisual pedagogy, where producers systematically developed multimodal techniques to teach children to align hearing and seeing. By studying the very audiovisual content, it is possible to explore the "television generation" was introduced to the "multimodality of communication". It was a practice that negotiated attention, and comprehension. Methodologically, this work shows how computational analysis can help trace historical patterns of audiovisual learning across archives, complementing close readings with quantitative sensitivity to rhythm and form of cross-modal aesthetics (Arnold & Tilton, 2019; Guldi, 2023).

At a time when multimodal AI systems are promoted with increasing conviction, it is crucial to understand how the very patterns these models learn from were historically produced. The experimental forms of 1970s children's television shaped the sensory habits of those who would later create the audiovisual culture of subsequent decades. By examining what these young audiences once consumed, we can better grasp the historical emergence of cross-modal logics, and, perhaps, glimpse alternative approaches to synchronization and meaning-making that risk being forgotten in an age of automated content production.

References

References

Altman, R. (1985). "The Evolution of Sound Technology." In E. Weis & J. Belton (Eds.), Film Sound: Theory and Practice. Columbia University Press.

Arnold, T., & Tilton, L. (2019). Distant Viewing: Computational Analysis of Film and Television.
Birdwhistell, R. L. (1970). Kinesics and Context: Essays on Body Motion Communication. University of Pennsylvania Press. http://www.jstor.org/stable/j.ctt3fhk73

Chion, M. (1994). Audio-Vision: Sound on Screen (C. Gorbman, Trans.). Columbia University Press.

Donnelly, K. (2019). The Spectre of Sound: Music in Film and Television. Bloomsbury.

Fickers, A., Snickars, P., & Williams, M. (2019). "Audiovisual Data in Digital Humanities." VIEW Journal of European Television History and Culture, 7(14).

Fiske, J. (1987). Television Culture. Methuen.

Guldi, J. (2023). The Dangerous Art of Text Mining: A Methodological History for the Digital Humanities. University of Chicago Press.

Hodge, R., & Tripp, D. (1986). Children and Television: A Semiotic Approach. Polity Press.
Janson, M. (2014). När bara den bästa TV:n var god nog åt barnen: Om sjuttiotalets svenska barnprogram. Stockholm: Karneval förlag.

Mares, M.-L., & Pan, Z. (2013). "Effects of Sesame Street: A Meta-Analysis of Children's Learning." Journal of Applied Developmental Psychology, 34(3).

Malmstedt, J. (2023). "Scale Exercises: Listening to the Sonic Diversity in 5000 Hours of Swedish Radio." Zoomland, 21.

Minow, N. (1961). "Television and the Public Interest." Speech to the National Association of Broadcasters.

Pecora, N. (2002). The Changing Nature of Children's Television: Fifty Years of Research. Lawrence Erlbaum.

Scannell, P. (1996). Radio, Television and Modern Life: A Phenomenological Approach. Blackwell.

Taggart, J., Eisen, S., & Lillard, A. S. (2019). The current landscape of US children's television: Violent, prosocial, educational, and fantastical content. Journal of Children and Media, 13(3), 276–294. https://doi.org/10.1080/17482798.2019.1605916

16:30–17:00 *LONG PAPER* **[74]**

Tracing Terrorism in Television: Toward a Methodology of Large-Scale Audiovisual Search

Daniel Brodén, Johan Malmstedt, Mats Fridlund

Gothenburg Research Infrastructure in Digital Humanities (GRIDH), Sweden

Keywords: *television history, distant viewing, search engines, terrorism*

1. Introduction

An iconic moment in Swedish television history is the 1975 broadcast from the West German embassy siege in Stockholm, a hostage stand-off initiated by the Red Army Faction (RAF). The dramatic climax, when a bomb explosion rocked the embassy building, has been etched into public memory through the grainy, night-time television footage of distressed news reporter Bo Holmström exclaiming: “Go live! Go live!” (“Lägg ut! Lägg ut!”). This moment, both live and discursively constructed, highlights television’s role in framing terrorism as a media event, while also marking the emergence of so-called international terrorism as an urgent national issue that has captured the Swedish public imagination since the late 1960s. In revisiting this moment through digitized archival broadcasts, we consider how such recordings circulate as shared cultural memory objects and how digital re-use reshapes access to Sweden’s broadcast heritage.

Although previous studies have shown that a modern notion of terrorism entered Swedish public discourse around 1970 (Fridlund et al. 2022; Brodén et al. 2023), the pivotal role of television in shaping its perception remains underexplored. This paper addresses this gap. Building on research that situates the modern notion of international terrorism as a discursive product of the global 1968 and the Cold War (Stampnitzky 2013; Zoller 2021), we develop and pilot a context-sensitive multimodal approach for large-scale analysis of historical television, specifically tracing how Swedish public service news broadcasting integrated “terrorism” as a modern media category.

Methodologically, we combine *distant viewing* (Arnold and Tilton 2019) with media historical contextualization (Guldi 2023), aligning speech-to-text outputs with visual and sonic features of news broadcasts, drawing on the digitized collections of the National Library of Sweden (*Kungliga biblioteket*; hereafter KB).

2. Digital Film and Television Studies

Our paper is situated within the emerging subfield of digital film studies (Heftberger 2019; Dang et al. 2024). As recently as 2019, observers noted that digital humanities methods remained largely “deaf and blind” to non-textual media (Fickers et al. 2018). Only recently have digital film studies scholars begun to use algorithms for image mining (Eriksson et al. 2022; Stelmach 2024), leaving television’s cross-modal character – where moving images, sound, and editing converge – underexplored. Over the past decade, there has been a gradual increase in scholarship engaging with speech-to-text methods for large-scale analysis and distant viewing of audiovisual data (Fickers et al. 2018, Heftberger 2019, Burghardt et al., 2020). Within the Swedish context, however, there are virtually no studies that systematically integrate broadcast audio and video as primary data (Stjernholm 2022).

Theoretically, we proceed from the assumption that aesthetic elements in television not only reflect but structure meaning-making. Because televisual narratives extend beyond mere representation of terrorist acts and public discourse, we understand broadcast mediation as shaped both semantically (statements, interviews, commentaries, voice-over, etc.) and formally (framing, editing, sound, etc.) (Prince 2011; Tavares Furtado and Auchter 2024). Accordingly, we treat television news as a site where period-specific and distinct repertoires of audiovisual narratives about terrorism are constructed.

3. Our Integrative Approach

This pilot consists of a large-scale multimodal analysis of television archival material, that traces (i) the lexical meanings attached to terrorism in speech, and (ii) how visuals and sound depict forms of political violence framed as terrorism, weighing these patterns against close readings and contextualization informed by television history and terrorism studies.

3.1 Sampled Broadcast Material

The data mining concerns Swedish public service television, which, until 1987, was the country’s sole provider of television news. In 1970, Sweden had a two-channel public service television system, with TV1 and TV2 both operated by Sveriges Radio (later Sveriges Television), offering complementary programming under a shared public broadcasting mandate.

KB houses extensive digitized collections of these broadcasts, enabling large-scale exploration. For this pilot, however, we focus on a randomized sample of 100 episodes of TV2’s flagship news program Rapport (launched in December 1969) from 1970–1979. As one of two daily national news programs in the period, alongside TV1’s *Aktuellt*, *Rapport* provides a stable lens for a preliminary examination of how public-service news represented terrorism as it entered Swedish discourse.

3.2 Multimodal Methodology

From the KB digitizations we build a time-aligned ASR corpus using an in-house Swedish Wav2Vec2 pipeline with light post-editing, yielding item-level segments suitable for search and concordance. On this text we derive monthly and yearly term series for the words *terror*, *terrorism*, *terrorist*, *kapning* (hijacking), *kidnappning* (kidnapping), *gisslan* (hostage), *gerilla*, *extremist*, and adjacent expressions, and then compute collocations, using simple change-point and burst detection to mark phases of lexical consolidation.

In parallel, we extract audiovisual features directly from the same news broadcasts: shot boundary and average shot length; proxies for shot scale; object and scene detections (masked figures, firearms, police cordons, embassy façades, airport tarmacs, emergency vehicles); and audio measures (speech/non-speech ratios, silence, sirens, explosions, crowd noise, microphone handling) together with prosodic markers of anchor and reporter urgency.

We then align the lexical and audiovisual timelines with lead-lag windows to test whether increases in montage density, non-speech loudness, and crisis iconography anticipate the stable naming of terrorism, and estimate uncertainty via resampling against year and item baselines. Flagged periods become cases for close reading of script, camera work, live inserts, and editing choices, keeping a distinction between terrorist and political violence as distinguishable events and terrorism as a journalistic label. The result is a transparent, replicable workflow confined to a single programme ecology that isolates how televisual form and vocabulary co-evolve.

3.3 Media Historical Contextualization

Balancing automated search with a media-historical approach that situates the broadcasts in their cultural context, we adopt a layered method that avoids narrow, purely data-driven analysis by attending to the “historical and contextual complexities of the documentary record” (Guldi 2023).

More broadly, television coverage of terrorism took shape alongside the expansion of television reporting on wars and conflicts around the world, the rise of live news, shifts in Swedish television aesthetics (including the adoption of color television), and the wider institutional and technological transformations of postwar public-service television (Weibull and Djerf-Pierre 2001; Bignell and Fickers 2008; Nykvist 2011; Djerf-Pierre and Ekström 2013; Scannell 2014; Wahlberg 2017).

Specifically, we move iteratively between distant viewing and close reading of the TV2 broadcasts to test and substantiate our results, noting relevant camera work, voice-over narration, and editorial structure. We also, to some extent, take into account the aestheticization of different forms of militant violence (aircraft hijackings, bombings, kidnappings, etc.) and situate the analysis within the wider historical contexts of the emergence of international terrorism in the 1970s, associated with militant organizations such as the PFLP, Black September, the RAF, and the IRA (Chouliaraki 2006; Melzer 2015).

4. Conclusions

We conclude by outlining how our context-sensitive multimodal approach yields methodological insights for large-scale analysis of macro-level patterns concerning the terrorism discourse in Swedish television news and for identifying specific segments suited to close reading. Although our integrative approach is exploratory, scaling it will produce a decade-long, single-programme corpus that is internally consistent enough for in-depth analysis yet remains representative of 1970s Swedish television news conventions. Perhaps most notably, preliminary findings from this limited case study suggest that visual and sonic features, to some extent, prefigures the subsequent stabilisation of the spoken discourse on modern terrorism discourse.

More broadly, the project demonstrates how digital methods can reactivate broadcast archives as sites of cultural memory, raising questions about curation, access, and the interpretive role of computational tools in shaping how the past is remembered.

References

Arnold, T, & L Tilton. 2019. “Distant Viewing: Analyzing Large Visual Corpora.” Digital Scholarship in the Humanities 34 (Issue Supplement_1): i3–i16.

Bignell, Jonathan, and Andreas Fickers, eds. 2008. A European Television History. John Wiley & Sons.

Brodén, Daniel, Leif-Jöran Olsson, Mats Fridlund, Magnus Ängsal, and Patrik Öhberg. 2023. “The Diachrony of the New Political Terrorism: Tracing Neologisms and Frequencies of Terror-related Terms in Swedish Parliamentary Data 1971–2018.” In DHNB 2022: Proceedings of the 7th Digital Humanities in the Nordic and Baltic Countries Conference (DHNB 2023), CEUR-WS: 79–89.

Burghardt, Manuel, Adelheid Heftberger, Johannes Pause, Niels-Oliver Walkowski, and Matthias Zeppelzauer. 2020. “Film and Video Analysis in the Digital Humanities – An Interdisciplinary Dialog. Digital Humanities Quarterly, 14 (4).

Chouliaraki, Lilie. 2006. Spectatorship of Suffering. Sage.

Dang, Sarah-Mai, Tim van der Heijden, and Christian Gosvig Olesen, eds. 2024. Doing Digital Film History: Concepts, Tools, Practices, De Gruyter.

Djerf-Pierre, Monika, and Mats Ekström, eds. 2013: A History of Swedish Broadcasting: Communicative Ethos, Genres and Institutional Change, Nordicom.

Eriksson, Maria, Tomas Skotare, Pelle Snickars. 2022. “Understanding Gardar Sahlberg with Neural Nets: On Algorithmic Reuse of the Swedish SF Archive’, Journal of Scandinavian Cinema 12 (3): 225–47.

Fickers, Andreas, et al. 2018. “Editorial: Special Issue Audiovisual Data in Digital Humanities.” VIEW: Journal of European Television History and Culture 7 (14): 1–4.

Fridlund, Mats, Daniel Brodén, Leif-Jöran Olsson, and Magnus P Ängsal. 2022: “Codifying the Debates of the Riksdag: Towards a Framework for Semi-automatic Annotation of Swedish Parliamentary Discourse”. In Proceedings of Digital Parliamentary Data in Action (DIPADA 2022) workshop at The 6th Digital Humanities in the Nordic and Baltic Countries Conference, edited by Matti La Mela, Fredrik Norén, and Eero Hyvönen. CEUR-WS: 167–75.

Guldi, Jo. 2023. The Dangerous Art of Text Mining: A Methodology for Digital History. Cambridge University Press.

Heftberger, Adelheid. 2019: Digital Humanities and Film Studies: Visualizing Dziga Vertov’s Work. Springer.

Melzer, Patricia. 2015. Death in Shape of a Young Girl: Women’s Political Violence in the Red Army Faction. New York University Press.

Nykvist, Ari. 2011. Formaterade nyheter: Studier i hur tv-nyheter formatmässigt editeras, gestaltas och tas emot, diss, Åbo Akademi University Press.

Prince, Stephen. 2011. Firestorm: American Film in the Age of Terrorism. Columbia University Press.

Scanell, Paddy. 2014. Television and the Meaning of “Live”. Wiley.

Stampnitzky, Lisa. 2013. Disciplining Terror: How Experts Invented “Terrorism”. Cambridge University Press.

Stelmach, Milosz. 2024. “Cinema Counts: The Computational Turn and Quantitative Methods in Film Studies.” Kwartalnik Filmowy (127): 6–28.

Stjernholm, Emil. 2022. “Distant Reading Televised Public Information: The Communication of Swedish Government Agencies, 1978–2020.” Necsus: European Journal of Media Studies 11 (2): 201–225.

Tavares Furtado, Henrique, and Jesse Auchter. 2024. “Demystifying Trauma in International Relations Theory: From Incomprehensibility to the Liberatory Real.” Security Dialogue 56 (1): 76–93.

Wahlberg, Malin. 2017. “Vietnam in Transmission.” Journal of Scandinavian Studies 7 (1): 43–64.

Weibull, Lennart, and Monika Djerf-Pierre. 2001. Spegla, granska, tolka: Aktualitetsjournalistik i svensk radio och TV under 1900-talet. Prisma.

Zoller, Silke. 2021: To Deter and Punish: Global Collaboration Against Terrorism in the 1970s. Columbia University Press.

17:00–17:20 *SHORT PAPER* **[75]**

Scene-Anchored Analysis of Affective “Stickiness” in Nordic Political-Thriller Television with Large Language Models

Aida Gholami

Leiden University, Netherlands, The

Keywords: *Affective stickiness, LLMs, stereotypes, embedding analysis, Nordic television*

Decades of research show that despite the abundance of training data, large language models (LLMs) internalize specific cultural regularities, including stereotypical associations and affective skews surrounding marginalized communities (Caliskan et al. 2017; Blodgett et al. 2020; May et al. 2019; Plaza-del-Arco et al. 2024). Recent research has shown that transformer-based models systematically attach particular emotions to social groups; associating anger, for instance, with Black identities (Kiritchenko and Mohammad 2018; Abid et al. 2021). While psychology often treats emotions as internal

states, in the humanities, affect theory emphasizes that emotions are social forces that circulate and "stick" to bodies, objects, and symbols (Brennan 2004; Ahmed 2014). This implies that in the extensiveness of digital discourse, the individual distinction of a marginalized identity is often lost and replaced by entities that embody collective emotions, for instance, when a woman wearing a hijab signifies threat or hostility.

Building on Abdel-Fadil's (2023a, 2023b) observation that media framings conflate categories like "Muslim" and "immigrant" into emotionally charged sites, the present study develops a methodology to examine whether LLMs reproduce this "affective stickiness" in narrative outputs. As a situated case study of genre-specific generation, we focus on narrative outputs within the high-contrast context of Nordic political thrillers (Caliphate 2020; Bullets 2018). Rather than eliciting refusals or safety-conditioned responses via zero-shot questioning, this analysis utilizes narrative completion to access distributions over plausible continuations where affective associations remain implicit.

The research relies on a controlled generation experiment designed to isolate the variable of identity. Minimal narrative scenes are constructed in three versions: targeted (Middle Eastern-identified), controlled (Nordic), and neutralized. We elicited [N=288] continuations from [GPT-4, Llama-3, and Mistral-Large], using controlled decoding parameters (temperature=0.7) to ensure reproducibility while allowing for the necessary narrative variance.

Methodologically, this paper operationalizes Ahmed's concept of "stickiness" by treating emotions not as internal states, but as circulating forces that adhere to bodies and objects in discourse. This generated corpus is analyzed through two complementary modes:

Mode 1: Qualitative affect coding (interpretive analysis) Because continuations are narrative and open-ended, the primary analysis is manual, single-coder interpretive analysis. For each continuation, salient carriers are categorized under four theory-driven headings: (i) bodies (especially racialized or gendered); (ii) symbols and cultural artifacts (e.g., hijab, bag); (iii) words and language (e.g., names, discursive labels); and (iv) practices and rituals. Each carrier is then coded using a fixed taxonomy of eight affective labels derived from Ahmed's framework: fear, suspicion, pity, pride, shame, respect, hostility, and empathy. To mitigate the risks of interpretive subjectivity, coding followed a strict codebook anchored in quoted evidence, with a reflexive focus on the coder's positionality regarding the source genre.

Mode 2: Semantic embedding analysis (corroboration) To scale these observations computationally, we employ the SBERT family of sentence encoders (Reimers and Gurevych 2019). We compute cosine similarities between manually extracted carrier mentions and affective seed wordlists tailored to the eight labels. Rather than serving as ground truth, this computational layer corroborates the qualitative findings by revealing parallel affective clustering in the semantic space. It demonstrates that two distinct systems of bias, namely the human reader's interpretation and the embedding model's pre-trained geometry, converge on the same negative associations.

Results indicate a marked divergence across both domestic and institutional beats. In institutional fragments, target continuations more frequently load fear and suspicion around the focal character than matched controls. At the level of carriers, items such as hijab and bag recurrently attract scrutiny (keywords: search, conceal, nervous). In family exchanges, target continuation shows more shame (being put on the spot) and less empathy under otherwise identical setups. Crucially, the computational analysis aligns with the manual coding: the semantic vector space of target narratives shifts significantly toward 'threat' and 'surveillance' even when explicit slurs are absent.

This paper offers a reproducible, scene-level audit of affective attachment. It demonstrates that in the abundance of genre-specific generative content, the "sticky" weight of the stereotype effectively erases the neutrality of the individual.

References

Abdel-Fadil, Mona. 2023a. "Triggers and Tropes: The Affective Manufacturing of Online Islamophobia." International Journal of Communication 17. https://ijoc.org/index.php/ijoc/article/view/13521.

Abdel-Fadil, Mona. 2023b. "Affective Witnessing of the Hijab: A Self-Inflicted Trauma." In The Routledge Companion to Gender and Affect, edited by Todd W. Reeser, 366–373. New York: Routledge.

Abid, Abubakar, Maheen Farooqi, and James Zou. 2021. "Persistent Anti-Muslim Bias in Large Language Models." In Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society, 298–306. New York: ACM. https://doi.org/10.1145/3461702.3462624.

Ahmed, Sara. 2014. The Cultural Politics of Emotion. 2nd ed. Edinburgh: Edinburgh University Press.

Behrman, Wilhelm, and Niklas Rockström, creators. 2020. Caliphate [TV series]. Stockholm: SVT / Netflix.

Blodgett, Su Lin, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. "Language (Technology) Is Power: A Critical Survey of 'Bias' in NLP." In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 5454–5476. Stroudsburg, PA: Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.485.

Brennan, Teresa. 2004. The Transmission of Affect. Ithaca, NY: Cornell University Press.

Caliskan, Aylin, Joanna J. Bryson, and Arvind Narayanan. 2017. "Semantics Derived Automatically from Language Corpora Contain Human-like Biases." Science 356 (6334): 183–186. https://doi.org/10.1126/science.aal4230.

Kiritchenko, Svetlana, and Saif M. Mohammad. 2018. "Examining Gender and Race Bias in Two Hundred Sentiment Analysis Systems." In Proceedings of the 7th Joint Conference on Lexical and Computational Semantics (SemEval-2018), 43–53. Stroudsburg, PA: Association for Computational Linguistics. https://doi.org/10.18653/v1/S18-2005.

May, Chandler, Alex Wang, Shikha Bordia, Samuel R. Bowman, and Rachel Rudinger. 2019. "On Measuring Social Biases in Sentence Encoders." In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics, 622–628. Stroudsburg, PA: Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1063.

Pesonen, Antti, and Minna Virtanen, creators. 2018. Bullets [TV series]. Helsinki: Elisa Viihde / MTV3.

Plaza-del-Arco, Flor Miriam, Alba Cercas Curry, Silvia Paoli, and Dirk Hovy. 2024. "Divine LLaMAs: Bias, Stereotypes, Stigmatization, and Emotion Representation of Religion in Large Language Models." Findings of EMNLP 2024, 3386–3405. https://aclanthology.org/2024.findings-emnlp.251/.

Reimers, Nils, and Iryna Gurevych. 2019. "Sentence-BERT: Sentence Embeddings Using Siamese BERT-Networks." In Proceedings of EMNLP 2019, 3982–3992. Stroudsburg, PA: Association for Computational Linguistics. https://doi.org/10.18653/v1/D19-1410.

17:20–17:40 *LONG PAPER* **[76]**

Framing Digital Transformation: Media Discourses on Digitalization in Sweden

Coppélie Cocq, Stefan Gelfgren, Rebecka Weegar

Umeå University, Sweden

Keywords: *Digital transformation; digitalisation; AI; text analysis*

In this paper, we will examine discourses of technological progress in Swedish news media. In times of fast digital transformation, this paper aims to interpret and deconstruct the process of the transformation, in its complexity by studying discourses and narratives about representations and expectations of digitalization.

Based on a distant reading of selected national news media, this paper will first examine how discourses about digitalization, digital transformation and AI have evolved during a period from the early 2000s to today. The results of this text analysis will be illustrated with the help of visualization tools (e.g. Word rains, Cirrus and Links) in order to highlight recurring patterns. Second, we will examine a selection of news articles from the same time period, in order to contextualize the content of these discourses, and the role of various actors in society in constructing, maintaining and challenging these discourses.

We focus on the Swedish context, a relevant case as it is a country with an ambition to be at the forefront of the digital transformation and with a long history of "datafying" its citizens through the welfare systems and bureaucracy. In Sweden, and the Nordic countries, the history of datafying its citizens has created a situation where data is seen as a "data gold mine" – giving the foundation and promises of a successful digital transformation (Tupasela et al. 2020).

Session 4D — 16:00–18:00

16:00–16:30 **Detecting Climate Delay Discourses in Danish Parliamentary Speeches: A Large Language Model Approach**
Camilla Buur Küseler, Florian Meier

16:30–17:00 **Decoding Diplomacy: A Large Language Model Approach to Emotional Rhetoric Across EU Foreign Policy Institutions**
Sasha Nielsen

17:00–17:20 **Tracing sentiment in the political discourse on homosexuality in the German Reichstag, 1895–1914**
Anna Maria Ramm

17:20–17:40 **From Low-Code to Open Code: Reflecting on a Developing Workflow for Analysis of Far-Right Discourse**
Daniel Ihrmark, Hanna Carlsson, Fredrik Hanell

17:40–18:00 **Carniolan Provincial Assembly: Corpus Improvements and Enhancements**
Ajda Pretnar Žagar, Kristina Pahor de Maiti Tekavčič

16:00–16:30 *LONG PAPER* **[77]**

Detecting Climate Delay Discourses in Danish Parliamentary Speeches: A Large Language Model Approach

Camilla Buur Küseler, Florian Meier
Aalborg University, Copenhagen, Denmark

Keywords: *climate change, parliamentary speeches, climate delay discourses, large language models*

As scientific consensus on anthropogenic climate change has solidified, outright denial in political debate has largely given way to more subtle forms of climate change rejection. These so-called climate delay discourses acknowledge the problem yet justify inaction, posing a growing challenge to effective climate communication and policymaking. This study operationalises Lamb et al.'s (2020) typology of climate delay discourses to examine 25 years of Danish parliamentary speeches (1998–2022). Using a keyword-based retrieval method, we identified approximately 34,000 climate change-related speech segments and applied large language models (LLMs) to classify them according to delay discourse categories. We compared zero-shot and few-shot prompting strategies, including variations with chain-of-thought reasoning, to evaluate LLM performance on complex rhetorical classification tasks. Few-shot prompting delivered promising results with respect to both recall and accuracy, while chain-of-thought reasoning provided limited benefits and, in some cases, harmed performance. Temporal and partisan analyses reveal that delay discourses have been consistently present in Danish political debate, with a marked increase in recent years. The most prevalent discourse, *all talk, little action*, reflects the gap between ambitious climate targets and policy implementation, particularly among governing parties. While right-leaning parties often shift responsibility away from Denmark, left-leaning and green parties more frequently invoke appeals to social justice. Our findings demonstrate both the promise and the limitations of LLMs for large-scale political discourse analysis and provide evidence that climate-delay discourses are a routine part of Danish parliamentary debates.

16:30–17:00 *LONG PAPER* **[78]**

Decoding Diplomacy: A Large Language Model Approach to Emotional Rhetoric Across EU Foreign Policy Institutions

Sasha Nielsen
Aarhus University, Denmark

Keywords: *EmoRoBERTa, sentiment analysis, EU foreign policy, emotional rhetoric, digital humanities*

This paper presents a methodological and empirical contribution to both international relations (IR) and digital humanities by applying a large language model (LLM) to investigate emotional rhetoric across various European Union foreign policy (EUFP) institutions. As part of a PhD project on the role of emotions in EUFP narratives, this study isolates the EU community's institutional dimension to examine how emotional expression varies across five key EUFP institutions: the European Council (EC), European Commission (EComm), Foreign Affairs Council (FAC), High Representative (HR), and the European External Action Service (EEAS). The central research question asks: *How are emotional*

rhetoric patterns differentiated across EUFP institutions, and in what ways do these rhetorical differences reflect and reinforce the institutions' respective mandates?

The paper engages with the DHNB 2026 theme by uncovering latent emotional structures within a vast corpus of institutional texts – materials that are abundant yet rarely examined for their affective content. EU foreign policy statements, often dismissed as being formulaic and procedural due to their diplomatic nature (see Smith, 2021; Keukeleire et al., 2016; Jørgensen, 1999), are here re-interpreted as emotionally-expressive resources that reflect institutional identity, strategic communication, and collective sentiment. By redirecting attention from traditional diplomatic analyses towards the emotional undercurrents present in the analysed rhetoric, this study contributes to an ever-growing body of work that seeks to reframe the 'challenge' of tracing emotions as an opportunity for deeper insights into foreign policy analysis (Gürkan and Terzi, 2024; Smith, 2021; Clément and Sangar, 2018; Bleiker and Hutchinson, 2018).

To address the research question, the paper employs EmoRoBERTa, a fine-tuned sentiment analysis model built on RoBERTa (Liu et al., 2019) and trained on the GoEmotions dataset (Demszky et al., 2020)[1]. EmoRoBERTa detects 27 distinct emotions in text-based data, enabling a granular analysis of emotional rhetoric. The model is applied to a curated corpus of official EUFP statements from 1 January, 2014 to 31 December, 2025, separated by the five institutional affiliations stated above. The current corpus includes over 37,000 filtered documents in total.

To guide its analytical framework, the paper utilises the emotional catalyst analytical structure introduced by Nielsen (2025), distinguishing between the three levels of emotional expression:

- Level 1 captures initial institutional responses to emotional catalysts, identifying short-term emotional spikes and rhetorical shifts.
- Level 2 examines long-term emotional patterns, using change point detection algorithms (BinSeg and PELT) to trace where sustained shifts occur in emotional rhetoric.
- Level 3 assesses the institutionalisation of emotion in rhetoric, employing multivariate statistical techniques (PERMANOVA, cosine similarity, hierarchical clustering, and PCA) to compare overall emotional profiles across the institutions.

The Level 1 analysis reveals distinct emotional trajectories across the institutions. For the EC, the top increasing emotions are desire, pride, and confusion, while gratitude, admiration, and grief decline. EComm's top increasing emotions are gratitude, admiration, and caring, while the top decreasing emotions are optimism, realization, and grief. The FAC displays a different emotional mix: grief, annoyance, and anger increase, while realization, desire, and optimism decline. The HR also shows a unique emotional profile: disapproval, pride, and anger increase, while realization, joy, and desire decline. Lastly, the EEAS has curiosity, gratitude, and disapproval as its top increasing emotions, while realization, anger, and optimism were the top decreasing emotions. The repetition of some emotions, such as realization and desire prove interesting, while the individual nature of emotions like joy are also telling. Overall, the top 5 increasing emotions were gratitude, curiosity, disapproval, sadness, and confusion; while the top 5 decreasing emotions were realization, optimism, grief, joy, and excitement.

The Level 2 analysis focuses on sustained emotional shifts. Across the institutions, the amount of change points that correlate with each other for each of the top 5 increasing and decreasing emotions decreases over time. For instance, prior to ultimo 2019, there were overlaps of change points for emotions such as curiosity and sadness expressed by the institutions, but after this time period, no two institutions experienced the same change point. This signals the increasing development of individual emotional profiles by the institutions over the given time frame.

Lastly, the Level 3[2] analysis confirms this institutional differentiation; specifically, the cosine similarity and clustering analyses reveal distinct emotional profiles across most of the institutions. The PCA analysis shows that FAC and HR are positioned on opposite sides along PC2, suggesting contrasting emotional tendencies rather than alignment. The EC appears separated along PC1, indicating its unique emotional signature. The hierarchical clustering confirms these distinctions: FAC and EC form the most distant clusters, while EComm and EEAS are more closely linked, reflecting moderate similarity in their emotional patterns. Overall, these results show that EC emphasizes strategic emotions such as desire, the FAC and HR exhibit divergent emotional activations of positive and negative emotions, and EEAS and EComm share moderate similarity, reflecting a distinct but interconnected set of emotional dynamics.

Taken together, these results suggest that emotional rhetoric in EU foreign policy statements is *indeed differentiated* among the EUFP institutions, but is *functionally aligned with their institutional mandates*. The EC continues to act as a political signaller, deploying emotions strategically to frame unity and rupture; the FAC communicates as the executor of consensus through gradually layered emotional cues such as annoyance and anger; and the HR adopts a personalised, humanistic emotional register consistent with its diplomatic role. Moreover, EComm exhibits a communicative style grounded in reassurance and empathic alignment, while the EEAS combines curiosity-driven framing with calibrated expressions reflective of its external-relations mandate. Crucially, the analyses demonstrate that these emotional dynamics are not peripheral rhetorical embellishments but integral components of each institution's communicative mandate, showing that EUFP statements routinely use emotional rhetoric to articulate and reinforce their intended institutionalised narratives.

In all, the applied analytical framework makes it possible to identify distinct emotional profiles and assess their coherence, as well as trajectory, within EU strategic communications. The analysis shows that emotional expressions are not marginal but structurally reinforced across institutions, with each of the five institutions exhibiting differentiated, sustained emotional patterns. Specifically, the emotional catalyst analytical structure provides a framework for tracing how emotions in rhetoric can develop into institutionalised narratives. This approach aligns with both constructivist and cognitivist theories, which emphasise the role of emotions in norm formation, identity construction, and policy signalling, where institutions deploy emotion rhetoric to signal unity and cultivate specific responses as part of strategic communicative 'practices' (Averill, 1980; Hoffmann, 2017; Koschut, 2018; Mercer, 2014; Aggestam, 2004).

Methodologically, by applying EmoRoBERTa to a non-canonical corpus of foreign policy texts, this paper showcases how digital methods can uncover hidden emotional architectures in political discourse. EmoRoBERTa's ability to capture nuanced emotional language highlights the transformative potential of LLMs for previously 'un-measurable' aspects of research within the humanities, such as emotions. Notably however, the paper also reflects on the methodological challenges associated with these new methods, including interpretability biases and training data limitations – as well as EmoRoBERTa's inability to show specifically where it detects emotion in texts. Finally, because the analysis does not examine the European public's reception, further research is needed to evaluate how such institutional narratives are interpreted by different levels of EU actors. Ultimately, models such as EmoRoBERTa offer computationally innovative and conceptually creative ways of rethinking the study of foreign policy in IR and the digital humanities.

[1] The code for the paper, including the data scraping, corpus, and EmoRoBERTa results, will be available on GitHub following publication. Until then, it can be obtained upon request from the author via sasni@cas.au.dk.

[2] The PERMANOVA results (pseudo-F = 201.965, p = 0.001) indicate statistically significant differences in emotional profiles.

References

Aggestam L (2004) Role Identity and the Europeanisation of Foreign Policy: A Political-Cultural Approach. In: Tonra B and Christiansen T (eds) Rethinking European Union Foreign Policy. Manchester: Manchester University Press, pp.81-98.

Averill J (1980) A Constructivist View of Emotion. In: Plutchik R and Kellerman H (eds) Emotion: Theory, Research and Experience: Vol. I. Theories of Emotion. New York: Academic Press.

Bleiker R and Hutchinson E (2018) Methods and Methodologies for the Study of Emotions in World Politic. In: Clément M and Sangar E (eds) Researching Emotions in International Relations. Palgrave Studies in International Relations.

Clément M and Sangar E (2018) Researching Emotions in International Relations: Methodological Perspectives on the Emotional Turn. Palgrave MacMillan.

Demszky D, Movshovitz-Attias D, Ko J, et al. (2020) GoEmotions: A Dataset of Fine-Grained Emotions. Association for Computational Linguistics. 4040-4054.

Gürkan S and Terzi Ö (2024) Emotions in EU foreign policy - when and how do they matter? Journal of European Integration 46(5): 575-596.

Hoffmann MJ (2017) Norms and Social Constructivism in International Relations. Oxford Research Encyclopedia of International Studies.

Jørgensen KE (1999) Modern European Diplomacy: A Research Agenda. Journal of International Relations and Development 2(1).

Keukeleire S, Keuleers F and Raube K (2016) The EU, Structural Diplomacy and the Challenge of Learning. In: Smith M, Keukeleire S and Vanhoonacker S (eds) The Diplomatic System of the European Union. Evolution, Change and Challenges. New York: Routledge, pp.199-214.

Koschut S (2018) The Power of (Emotion) Words: On the Importance of Emotions for Social Constructivist Discourse Analysis in IR. Journal of International Relations and Development 21(3): 495-522.

Liu Y, Ott M, Goyal N, et al. (2019) RoBERTa: A Robustly Optimized BERT Pretraining Approach. DOI: https://doi.org/10.48550/arXiv.1907.11692.

Mercer J (2014) Feeling Like a State: Social Emotion and Identity. International Theory 6(3): 515-535.

Nielsen SJ (2025) Rhetorical Shifts in European Union Foreign Policy: The Impact of Emotional Catalysts Following the Full-Scale Invasion of Ukraine. [Unpublished Manuscript].

Smith KE (2021) Emotions and EU foreign policy. International Affairs 97(2).

17:00–17:20 *SHORT PAPER* **[79]**

Tracing sentiment in the political discourse on homosexuality in the German Reichstag, 1895–1914

Anna Maria Ramm

University of Vienna, Austria

Keywords: *Sentiment Analysis, Gender History, Tool Criticism, Discourse Analysis, Reichstagsprotokolle*

This paper presents a case study in sentiment analysis applied to an historical corpus of parliamentary debates, exploring how digital methods can uncover marginalised discourses within canonical political language – marginal in terms of their low frequency in parliamentary debates. Situated at the intersection of gender history, discourse analysis and digital humanities, the study investigates how male homosexuality was discussed or alluded to in the German *Reichstag* between 1895 and 1914. Although the parliamentary protocols have long served as a source for political and institutional history, this project approaches them as a site of emotional and ideological negotiation, using computational methods to reveal patterns that remain invisible to traditional hermeneutic reading.

Homosexuality, in particular, was a topic whose explicit articulation was constrained by moral, legal and rhetorical conventions at the turn of the century. Consequently, traces of this discourse are fragmented and often encoded through euphemisms and indirect references. The paper argues that digital methods allow us to trace these faint signals across large textual corpora and to reassess how attitudes and emotions were linguistically encoded in political speech.

Building on this, the study is guided by a two-part research question. First, on a substantive level, it examines the long-standing historiographical claim that the Eulenburg Affair (1906–1909) intensified negative and more strongly moralised debates on homosexuality in the German *Reichstag*. It asks whether quantitative methods can support or challenge this assumption. Second, the study focuses on the methodological question of how the discourse can be operationalised: how can relevant passages be located through keyword-based retrieval and how can the accuracy and limitations of the sentiment analysis be critically evaluated?

The empirical basis of this study is a self-compiled dataset derived from the digital *Reichstagsprotokolle* collection, encompassing speeches and complementary documents from 1867 to 1942. From over 290,000 digitised pages, 118 relevant entries were identified using an historically informed keyword strategy. Because the term ‘homosexuality’ was only gradually adopted, a broader range of search terms was used to create the corpus. The selection of search terms was derived both from historiographical literature and from contemporary discursive fields such as law and medicine as well as references to public figures and institutions. Multiple historical spellings and orthographic variants were included to account for linguistic change and OCR-related distortions in the digitised material. Given the scale of the corpus and the prevalence of indirect or ‘hidden’ references, the study deliberately refrained from using topic modelling or automated text-discovery approaches and instead relied on a carefully curated keyword strategy.

Methodically, the study applies a lexicon-based sentiment analysis, supplemented by contextual analyses of co-occurring terms and frequencies. The German-language dictionary *SentimentWortschatz* (SentiWS for short) by Robert Remus, Uwe Quasthoff and Gerhard Heyer was selected for the study since it does not only include adjectives and adverbs but also nouns and verbs which allows for a broader capture of evaluative language. Previous research described the dictionary as promising in terms of analytical accuracy, while also cautioning against the risk of false positives when applied to historical texts. This limitation is considered in the present study. Preprocessing involved normalisation of historical spelling, correcting OCR errors and removing stopwords to ensure comparability. The analysis distinguishes between two levels of emotional expression: On the one hand, overall sentiment scores of complete speeches and on the other hand, local sentiment in the immediate context surrounding the search terms. By comparing these levels, the study aims to capture not only explicit statements but also the implicit tonal shifts that accompany mentions of homosexuality.

The results indicate distinct emotional patterns in the *Reichstag's* discourse. References to homosexuality increase during the Eulenburg Affair, yet the tone of the debates does not become more negative. The analysis thus challenges the assumption that the affair fundamentally intensified negative discourse. Rather, differences in sentiment follow political alignments more than the topic itself, suggesting that emotional language operated as a rhetorical resource within parliamentary debate.

The *Reichstagsprotokolle* represent a highly canonical corpus. Yet, through digital text analysis, it becomes possible to detect marginal topics and emotional undercurrents.

The application of sentiment analysis to historical sources also exposes methodological limits. As the model evaluates words in isolation, it cannot fully grasp rhetorical nuance, irony or contextual inversion. These challenges highlight the importance of tool criticism and the continued dialogue between quantitative output and qualitative reading.

By reframing the political discourse on homosexuality as an emotional discourse, this study demonstrates how digital humanities methods can expand the reach of historical inquiry. The combination of computational sentiment analysis and contextual interpretation opens a pathway toward understanding how emotions were mobilised to define the boundaries of inclusion and exclusion.

The case study argues for a more reflexive and critical use of digital methods in the humanities: not as tools of quantification alone, but as instruments for re-reading historical archives. The digital abundance of sources invites us to seek out what remains lost within it: the muted emotions, the marginalised subjects and the subtle textures of language that digital analysis can bring to light.

References

Bruns, Claudia: The Politics of Masculinity in the (Homo-)Sexual Discourse (1880 to 1920), in: German History 23 (2005) 3, pp. 306–320.

Domeier, Norman: The Homosexual Scare and the Masculinization of German Politics before World War I, in: Central European History 47 (2014) 4, pp. 737–759.

Freis, David: Homosexualität und Männlichkeit im Spannungsfeld von Justiz, Psychiatrie, Militär und Adel. Ein Fall aus der forensischen Militärpsychologie des Ersten Weltkriegs, in: kultur & geschlecht 7 (2011), pp.. 1–29.

Foucault, Michel: Sexualität und Wahrheit. Der Wille zum Wissen 1, übers. von Ulrich Raulf/Walter Seitter, Nördlingen 1977.

Haslinger, Peter: Diskurs, Sprache, Zeit, Identität. Plädoyer für eine erweiterte Diskursgeschichte, in: Eder, Franz X. (ed.): Historische Diskursanalyse. Genealogie, Theorie, Anwendungen, Wiesbaden 2006, pp. 27–50.

König, Mareike: Digitale Methoden in der Geschichtswissenschaft. Definitionen, Anwendungen, Herausforderungen, in: Zeitschrift für Biographieforschung, Oral History und Lebensverlaufsanalyse 30 (2017) 1–2, pp. 7–21.

Küsters, Anselm: Technological change and sentiment in German parliamentary speeches (1867–1942). An explanatory case study on telecommunication, in: Althage, Melanie et al. (edd.): Digitale Methoden in der geschichtswissenschaftlichen Praxis. Fachliche Transformationen und ihre epistemologischen Konsequenzen. Konferenzbeiträge der Digital History 2023, Berlin, 23.–26.05.2023, Berlin 2023, pp. 1–12.

Landwehr, Achim: Geschichte des Sagbaren. Einführung in die Historische Diskursanalyse, Tübingen 2001.
Nieden, Susanne zur: Homophobie und Staatsräson, in: Nieden, Susanne zur (ed.): Homosexualität und Staatsräson. Männlichkeit, Homophobie und Politik in Deutschland 1900–1945 (Geschichte und Geschlechter 46), Frankfurt/New York 2005, pp. 17–51.
Rauh, Christian: Validating a sentiment dictionary for German political language – a workbench note, in: Journal of Information Technology & Politics 15 (2018) 4, pp. 319–343.
Spugnoli, Rachele et al.: Towards sentiment analysis for historical texts, in: Digital Scholarship in the Humanities 31 (2016) 4, pp. 762–772.
Thaller, Manfred: Geschichte der Digital Humanities, in: Jannidis, Fotis/Kohle, Hubertus/Rehbein, Malte (edd.): Digital Humanities. Eine Einführung. Mit Abbildungen und Grafiken, Stuttgart 2017, S. 3–12.

17:20–17:40 *SHORT PAPER* **[80]**

From Low-Code to Open Code: Reflecting on a Developing Workflow for Analysis of Far-Right Discourse

Daniel Ihrmark, Hanna Carlsson, Fredrik Hanell
Linnaeus University, Sweden

Keywords: *sentiment analysis, topic modeling, dashboards, low-code, Python, cultural institutions*

1. Introduction

In Sweden, cultural institutions such as public libraries and museums have become politicized arenas within an emerging “culture war,” where the far-right mobilizes online to challenge their democratic mission (Harding, 2021; Carlsson et al., 2022). The Cultural Institutions and the Culture War (CiCuW) project analyzes far-right discourse across online platforms to examine how ideological framings emerge and circulate. Building on the project’s earlier low-code pilot using LDA topic modelling and zero-shot sentiment analysis based on RoBERTa (Hanell et al., 2025), this paper presents a fully open, reproducible Python workflow for topic modeling, sentiment analysis, and zero-shot classification, implemented in a Streamlit web application. It reflects on how the shift from low-code to scripted workflows reshapes the project’s methodology and its pedagogical outputs in relation to current discussions in Digital Humanities education (Goicoechea et al., 2025).

2. From Low-Code Pilot to Python Project

The CiCuW pilot, published in the *Huminfra Handbook* (Ihrmark, Carlsson & Hanell, 2025), used KNIME to perform sentiment analysis, topic modeling, and interactive visualizations. This phase demonstrated how researchers with limited programming experience could apply zero-shot BERT sentiment classification and LDA topic modeling within a visual environment (Ihrmark & Tyrkkö, 2023). However, limitations became evident: reliance on proprietary interfaces, reduced control over tokenization and parameters, and restricted portability (Tyrkkö & Ihrmark, 2024).

The current iteration reconstructs the analytical logic in an open, scripted environment enabling versioning, parameter transparency, and modular reuse. Low-code tools are reframed as pedagogical entry points toward open workflows. This aligns with Goicoechea et al.’s (2025, p. 8) emphasis on programming, data management, and AI competencies combined with critical understanding. The transition introduces conceptual shifts concerning availability, transparency, and abstraction, while positioning DH graduates as humanists who “understand technology and its culture” and “think programmatically and critically” (2025, p. 8).

3. The Data and the Workflow

The data for the project consists of online materials representing different registers and formats. The project’s focus on far-right discourse surrounding cultural institutions served as the main guideline for inclusion:

• YouTube: 3,571 transcribed videos (11 channels) with connections to the Swedish far-right, transcribed via Whisper Turbo and enriched with metadata from the YouTube API.

• Flashback: 6,638 forum posts from the Culture and Politics sub-forum, retrieved from Språkbanken Text (2025).

The pipeline comprises four stages designed to produce a dataset enabling comparative exploration of how topics connected to public institutions are discussed across mediated (YouTube) and participatory (forums) environments. Each stage can be reproduced independently or substituted with alternative models if issues with the current selection appear. In addition, the rapid development of Swedish language NLP resources is likely to result in better options for topic modelling and sentiment analysis becoming available before the project concludes. However, as the low-code workflow relies on extensions for KNIME being available, the quicker uptake for the Python workflow highlights one of the main benefits of the move. The Python workflow is presented below:

Stage 1: Topic modeling

BERTopic (Grootendorst, 2022) uses transformer embeddings to identify semantically coherent clusters across multilingual and noisy data. Dimensionality reduction and density clustering yield interpretable topics, which are then manually interpreted and labeled by the project participants. The initial topics are visualized interactively through the Streamlit dashboard.

Stage 2: Sentiment analysis

Polarity classification is applied at both document and topic levels using the NLPTown multilingual BERT (Peirsman, 2020). This enables the platform visualizations to show intersections of topics and sentiment, narrowing the scope of documents for qualitative analysis.

Stage 3: Zero-shot classification

Using user-defined labels, the inclusion of the KB-Lab Megatron BERT zero-shot model enables rapid thematic scoping of the dataset as well as testing of possible topic classifications. In addition, the zero-shot approach is also being considered for sentiment analysis. This is also the only interactive model included in the dashboard, while both topic modelling and sentiment analysis are carried out during preprocessing of the data.

Stage 4: Visualization (Streamlit Dashboard)

A Streamlit dashboard integrates BERTopic visualizations, sentiment overlays, and zero-shot classification outputs in an interactive interface. It supports comparisons and filtering. Metadata (such as publication date, platform, channel, and forum) is preserved for filtering and temporal analysis, allowing the project participants to narrow down the visualization to a specific selection.

Ethical engagement and methodological literacy remain central for the workflows. The workflow integrates principles from AoIR 3.0 (franzke et al., 2020) and computational thinking (Wing, 2006), maintaining the reflective ethos of the earlier low-code approach. The scripted environment preserves this accessibility while providing full transparency of methods, parameters, and model behavior. In addition, the public hosting of the code on GitHub highlights one of the conceptual shifts, as the project methodology moves towards becoming more openly available and reusable.

This design also serves a pedagogical function: the KNIME workflow continues to be used for training and demonstration, while the openly available Python implementation functions as an opportunity for scaling and customization. Together, they form a continuum from exploration to application. However, that continuum is still reliant on acquiring Python literacy in order to make the jump between workflow formats.

4. Conclusions

Transitioning from KNIME to an open Python workflow has allowed the *CiCuW* project to increase transparency, flexibility, and analytical depth while retaining accessibility for non-technical collaborators through its continued pedagogical use of low-code tools and through the Streamlit dashboard. The combined workflows illustrate how methodological literacy can evolve through iterative tool adoption. In addition, while the needs of the research project has driven the methodological component towards the Python implementation, the use of a low-code workflow for the initial formulation mimics the need for

methodological grounding and concretization needed at the early stages of educational programmes in order for students to initially understand the techniques involved through hands-on experimentation, before optionally working towards implementation through programming. Regardless of whether programming is pursued, the workflow of the low-code implementation still incorporates computational thinking as a pedagogical tool and programmatic thinking in the context of humanities projects, while forcing engagement with the ethical and critical considerations involved in applying such methods.

References

Carlsson, H., Hanell, F., & Hansson, J. (2022). ”Det känns som att jag bara sitter och väntar på att det ska explodera”: Politisk påverkan på de kommunala folkbibliotekens verksamhet. Nordic Journal of Library and Information Studies, 3(1), 26–43.

franzke, A. S., Bechmann, A., Zimmer, M., Ess, C., & AoIR (2020). Internet Research: Ethical Guidelines 3.0. Association of Internet Researchers.

Grootendorst, M. (2022). BERTopic: Neural topic modeling with a class-based TF-IDF procedure. arXiv.

Goicoechea, M., Sanz Cabrerizo, A. D. R., van der Lek, I., Gheldof, T., Joffres, A.,

Garnett, V., Weis, J., Kurzmeier, M., & Antloga, Š. (2025). Bridging the gap between industry and education: Best practices for the digital humanities (White paper). DARIAH. Zenodo. https://doi.org/10.5281/zenodo.17152979

Hanell, F., Carlsson, H., & Ihrmark, D. (2025). Exploring culture war related attacks on public libraries: Results from a pilot study. Information Research, 30(CoLIS), 344–365.

Harding, T. (2021). Culture wars? The (re)politicization of Swedish cultural policy. Cultural Trends, 30(1), 1–18.

Ihrmark, D., Carlsson, H., Hanell, F. (2025). Low-code web scraping and text analysis with Octoparse and KNIME : An example from the CICuW project. Huminfra Handbook: Empowering digital and experimental humanities. Tartu, University of Tartu. 505-540

Ihrmark, D., & Tyrkkö, J. (2023). Learning text analytics without coding? An introduction to KNIME. Education for Information, 39(2), 121–137.

Peirsman, Y. (2020). nlptown/bert-base-multilingual-uncased-sentiment [Computer software]. Hugging Face.

Sikora, J. (2023). The KBLab Blog: Swedish zero-shot classification model. KB Lab.

Tyrkkö, J., & Ihrmark, D. (2024). Low-code data science tools for linguistics: Swiss army knives or pretty black boxes? In S. Coats & V. Laippala (Eds.), Linguistics Across Disciplinary Borders (pp. 40–66). Bloomsbury.

Språkbanken Text. (2025). Flashback. https://doi.org/10.23695/YKK8-7D22

Wing, J. M. (2006). Computational thinking. Communications of the ACM, 49(3), 33–35.

17:40–18:00 *SHORT PAPER* **[81]**

Carniolan Provincial Assembly: Corpus Improvements and Enhancements

Ajda Pretnar Žagar[1,2], Kristina Pahor de Maiti Tekavčič[1,3]

[1] *Institute for Contemporary History, Slovenia*

[2] *Faculty of Computer and Information Science, University of Ljubljana*

[3] *Faculty of Arts, University of Ljubljana*

Keywords: *historical parliamentary proceedings, OCR correction, error analysis, metadata enrichment*

Historical parliamentary corpora offer crucial evidence for studying political discourse over time, yet their usability is often limited by poor OCR quality and incomplete metadata. This paper presents the enhancement of the Kranjska 1.0 corpus, a collection of Carniolan Provincial Assembly proceedings (1861–1913) in Slovene and German, through a two-phase process aimed at improving textual accuracy and enriching speaker metadata. First, we conducted a manual correction campaign on a representative sample of transcripts, involving trained historians proficient in Gothic script and 19th-century politics. The corrections addressed both structural and textual errors in TEI-encoded XML files, providing a gold-standard dataset for future model training. Error analysis revealed recurring OCR issues, including segmentation problems, misattributed speakers, and systematic character-level noise. Second, we harmonised and expanded speaker metadata using multiple historical sources to unify name variants, resolve ambiguities, and document parliamentary terms, factions, and attendance. The resulting metadata enhance corpus usability and interpretability. This work lays the foundation for the next project phase, which explores the automatic correction of transcripts and metadata using Multimodal Large Language Models (MLLMs). By combining historical expertise with computational

methods, we contribute to more accurate processing of historical texts and promote transparency and reusability in digital humanities research.

FRIDAY, 13 MARCH 2026

DT 1 — 08:30–10:00

08:30–09:00 **The odd one out: Conceptualizing and digitizing ephemeral jobbing prints**
Holger Berg

09:00–09:30 **How digital methods are taught: A mixed-method analysis of educational resources in spatial humanities**
Ahmad Kamal, Daniel Ihrmark, Patrick Gavin

09:30–10:00 **Assessing AI Recognition of Text and Symbols in Early Modern Cartographic Material**
Thomas Holgersson, Daniel Ihrmark, Henrik Svensson, Jonas Svensson, Ahmad Kamal

08:30–09:00 *DEVELOPMENTAL TRACK* **[82]**

The odd one out: Conceptualizing and digitizing ephemeral jobbing prints

Holger Berg
The Hans Christian Andersen Center, University of Southern Denmark, Odense

Keywords: *bibliography; mass digitization; ephemeral prints; GLAM; crowdsourcing*

1. Research questions

Next week's offers in the local supermarket. A flyer calling for climate action now. The song written to celebrate your aunt turning seventy. – Such shortlived prints have been archived systematically in the Nordic countries for well over a century. Anglophone research librarians speak of ephemeral prints and have begun incorporating such to diversify textual heritage (e.g., Waugh 2022).

What are the basic traits shared by this diverse textual mass? How can scholars explore these vast, scarcely catalogued holdings? My paper discusses how to define and digitize ephemeral prints, drawing on a recent case study where I manually catalogued 421 items (Berg 2023). This barely amounts to 1 of the 6.300 meters of prints shelved in Copenhagen, but the study showed their value for studying how literary texts circulated prior to commercial publication.

2. Grey, minor, mundane, ephemeral, occasional or incidental?

What are we to call the third type of printed texts, besides books and periodicals? Librarians apply overlapping adjectives: Many of these prints are as short-lived as a mayfly (i.e. ephemeral) and are at times used for specific festive occasions or daily purposes (the mundane hverdagstryk). Prints are frequently shorter than books (minor prints, småtryk) and were produced as incidental jobbing prints (accidenstryk). The latter term was used by contemporaries and is arguably most fitting: Unlike books, jobbing prints were produced at short notice and circulated outside book trade. The publishers active in this grey zone are rarely specialised in circulating such prints (Farace & Scköpfel 2010). While the printer shop is mentioned, a publishing house is generally absent.

Whatever the name, the cataloguing problems remain the same for libraries with large holdings. From 1902 to 1927 collections in the Royal Library increased tenfold. Storage was the main concern, not indexing. Jobbing prints were catalogued by less prominent librarians, often women, who fought hard to assert their importance.

3. The present indexing of collections by issuer and subject matter

Book publications have been catalogued for centuries and digitized for decades. More recently Nordic newspapers have also been indexed in detail, with entries on each issue. Lise Hesselager's 1974 article on Danish jobbing prints is the single overview of the much more general indexing first by subject matter and then by issuer. Since 1902, The Royal Library has systematically been collecting jobbing prints published in the Kingdom of Denmark, including domestic prints in German, Kalaallisut, Faroese and other languages. Jobbing prints published abroad are indexed even more sparsely.

Registration mostly remains manual. Subcollection headings have been transferred to the digital OPAC-catalogues, making each collection findable. More recently, they can be digitized on demand under a

70-year copyright-rule whose relevance should not be overstated (printed minutes from an annual assembly scarcely count as creative works).

By comparison Norwegian digitization aims at the entire cultural heritage held in documents, yet the outcome is far from complete and focuses on facsimiles and OCR, not metadata (Boasen et al 2024). Among Nordic national libraries only the Finns have, so far, embarked on a fine-grained registration of printed manifestations rather than the subcollections and the containers (boxes, bound codices, etc.).

The present indexing fits the corpus approach popular in DH-studies. Subcollections hold a ready-built corpus with known patterns of circulation. Researchers can instead concentrate on indexing and exploration.

4. How best to digitize jobbing prints and balance scale with rich metadata?

The present and future indexing tasks can be outlined via the FAIR acronym. Jobbing prints no longer must be consulted in reading rooms. Growing numbers are accessible as digital facsimiles with OCR text. However, these diverse prints will become far more findable if properly indexed. I propose two steps:

1) Research librarians should first digitize existing metadata in an interoperable format and add basic metadata by automated means.

2) This mix of verified and unverified metadata should be made available for reuse and refinement in smaller projects alongside image and text data.

This division of labor does justice to the giant task and research libraries' limited resources. Being unfamiliar with mass digitization best practices, I am particularly grateful for feedback on the following questions.

In step 1, which methods best address the problems peculiar to digitizing jobbing prints? They seem threefold:

a) To conceptually separate the visually identical items produced during the same print run from textual similarity between versions in separate imprints. The International Federation of Library Associations and Institutions advice its members to use maps relations between a work (the ideal text) and its expressions (text versions) and manifestations (print runs, documents) that are in turn each preserved in multiple physical items (copies, records; Riva, Le Bœuf & Žumer 2017; the FRBR-approach). Libraries hold multiple items of the same jobbing print. The items have deliberately been distributed in multiple subcollections.

b) How can such duplicate items be identified computationally? Aggregated digital collections like HathiTrust and Europeana have addressed the issue by comparing visual and textual data with metadata, but the clustering approaches known to me may now be superseded (e.g., Wang et al. 2013)

Existing analogue metadata is tailored to the given subcollection. Can data be digitized in all its richness and then incorporated into the general OPAC system through the simplified Dublin Core dot-annotation approach?For step 2, I would welcome best practices on:

a) Which controlled vocabulary is most suitable? Jobbing prints are structurally more diverse than books and periodicals. Should one use the broad Wikidata categories (currently 12,938) or alternative vocabularies?

b) Which platform setups currently function best? Image storage is probably best kept at the holding institution. What light-weight, extensible interfaces for collaborative bibliographical entry are most effective and user-friendly?

Could the enriched data be stored in long-term in the Zenodo repository and links added to the subcollection entry in the OPAC catalogues of the holding institution?

References

Berg, Holger (2023) “Increasing access to ephemeral prints: How to construct and analyze a dataset from the Golden Age of literature in nineteenth-century Denmark”, Orbis Litterarum, Volume 78, Issue 5, pp. 464–482.

Boasson, Frode Lerum; Malvik, Anders Skare; Eliassen, Knut Ove (2024). ”Den digitale litteraturarven og myten om ’alt’”, Nytt Norsk Tidsskrift, Volume 41(3–4), pp. 246–256.

Farace, Dominic; Schöpfel, Joachim, eds. (2010). Grey Literature in Library and Information Studies. Göttingen: De Gruyter/Saur.

Riva, Pat; Le Bœuf, Patrick; Žumer, Maja (August 2017). IFLA Library Reference Model: A Conceptual Model for Bibliographic Information. International Federation of Library Associations and Institutions. https://www.ifla.org/wp-content/uploads/2019/05/assets/cataloguing/frbr-lrm/ifla-lrm-august-2017.pdf

Hesselager, Lise (1974) ”Danske småtryk i Det kgl. Bibliotek”, Fortid og Nutid, Volume 24, pp. 63–98.

Wang, Shenghui; Isaac, Antoine; Charles, Valentine; Koopman, Rob; Agoropoulou, Anthi; van der Werf, Titia; Aalberg, Trond; Farrugia, Charles J.; Tsakonas, Giannis; Papatheodorou, Christos; Dobreva, Milen (2013) “Hierarchical structuring of Cultural Heritage objects within large aggregations” Research and Advanced Technology for Digital Libraries, 2013, pp. 247-259. Accessed via https://arxiv.org/pdf/1306.2866.

Waugh, Audrey (2022) “Alexander Turnbull Library Collecting Plan. Ephemera Collection” https://natlib.govt.nz/files/strategy/atl-ephemera-collecting-plan-updated-2022.pdf.

09:00–09:30 *DEVELOPMENTAL TRACK* **[83]**

How digital methods are taught: A mixed-method analysis of educational resources in spatial humanities

Ahmad Kamal[1], Daniel Ihrmark[1], Patrick Gavin[2]

[1] *Linnaeus University, Sweden*

[2] *Huron University, Canada*

Keywords: *instruction, online educational resources, spatial humanities, mixed-methods*

With the growing availability of free and open-source applications (e.g. Gephi, Python scripts, QGIS) and browser-based tools (e.g. Voyant Suite, OpenRefine, Omeka.Net, Flourish) access to digital technologies is no longer the primary barrier for humanities scholars. Instead, the challenge is discovering the right tools and developing the competencies needed to use them effectively. Fortunately, a wide range of online resources—such as YouTube tutorials, GitHub guides, and platforms like The Programming Historian—offer support for learning digital methods.

However, teaching digital methods in the humanities remains complex. Learners vary widely in their technical proficiency, disciplinary backgrounds, research needs, and conceptual understanding. As a result, educational resources reflect diverse pedagogical approaches and assumptions, which are often implicit and rarely articulated. Few studies have systematically examined these resources to understand their pedagogical positioning, making it difficult to match them to specific user groups or learning scenarios.

This work-in-progress addresses that gap by focusing on the case of spatial technology instruction in the digital humanities. Spatial technologies are frequently introduced in DH courses, textbooks (e.g. Drucker’s The Digital Humanities Coursebook, 2021), and library guides (e.g. NYU’s DH guide, https://guides.nyu.edu/digital-humanities). Yet outside specialized fields like archaeology or human geography, the use of geographic information systems (GIS) remains relatively rare. GIS tools often require a steep learning curve, technical expertise, and familiarity with complex data formats—posing a high barrier for scholars in non-technical disciplines. Moreover, many forms of humanistic inquiry are not naturally aligned with the literal spatiality demanded by GIS. Scholars such as Franco Moretti and Katherine Hayles have even argued that mapping may conflict with humanistic epistemologies, despite their potential. This dual challenge—technical and conceptual—makes spatial methods an ideal case study for exploring how digital tools are taught and understood in the humanities.

To investigate this, the on-going study analyzes thirty open educational resources (OERs) that introduce spatial technologies and techniques to humanists. These resources are drawn from prominent platforms: The Programming Historian (17), Dariah Campus (5), Esri’s GIS for Humanities (6), and DariahTeach (2). The goal is to understand how GIS technologies and spatial methods are presented to learners with primarily humanistic backgrounds.

The analysis combines manual thematic coding with computational topic modeling using BERTopic. Thematic analysis has thus far revealed six key attributes in digital method instruction:

Genre: format of the resource (e.g. walkthrough, cookbook, course module, project report)

Use-cases: the research question or problem used to demonstrate the tool's value

Structure: how the learning content is organized (e.g. step-by-step, essayistic, branching paths)

Technical concepts: explanations of GIS-related ideas (e.g. georeferencing, vector data, integrity testing)

Contextual concepts: non-technical ideas (e.g. spatial theory, gazetteers)

Audience: assumptions about users' background, proficiency, and motivations

Preliminary topic modeling reveals six clusters of OERs, broadly distinguishing resources by their technical orientation (platform-based vs. coding-based) and topical focus (GIS tools vs. spatial research).

These initial findings offer promising insights but also raise further questions. How might the study be enriched by integrating theories of learning or instructional design? Should the scope be expanded to include other document types, such as introductory texts on spatial approaches? Could the framework be applied to other digital methods, such as text analysis or network analysis?

By examining how spatial methods are taught to humanists, this study aims to illuminate broader patterns in digital humanities pedagogy and contribute to more effective, inclusive, and theoretically grounded instructional practices. We look forward to critical feedback to support us in this endeavour.

09:30–10:00 *DEVELOPMENTAL TRACK* **[84]**

Assessing AI Recognition of Text and Symbols in Early Modern Cartographic Material

Thomas Holgersson[1], Daniel Ihrmark[1], Henrik Svensson[2], Jonas Svensson[1], Ahmad Kamal[1]

[1] *Linnaeus University, Sweden*

[2] *Kristianstad University, Sweden*

Keywords: *historical maps, handwritten-text recognition, image recognition, artificial intelligence, performance evaluation*

Historical maps are rich cultural artifacts that pose significant challenges for computational analysis. Maps contain a complex assortment of textual and visual elements, including diverse annotations, non-standard layouts, and numerous symbolic representations. These features complicate the application of conventional optical character recognition (OCR) technologies. This presentation explores the feasibility of using contemporary artificial intelligence (AI) tools to identify handwritten placenames and topographic symbols—such as churches and fortifications—in early modern cartographic materials.

The project focuses on the unprinted works of Danish cartographer Johannes Mejer (1606–1674), whose maps can offer unique insight into 17th-century Nordic geography and culture; for instance, Mejer was among the first to chart the region of Skåne prior to its acquisition by Sweden. The corpus comprises over 200 digitized items from the Danish Royal Library, including maps, sketches, and handwritten documents. These copious materials has not been systematically transcribed or analysed thus far, presenting an opportunity to apply AI-based methods to unlock their content.

To evaluate the feasibility of currently available AI technologies for handwritten text recognition (HTR) and symbolic image recognition with respect to such historic maps, the project experiments with several AI-drive tools. Initial experiments have been conducted using Transkribus (German Giant model), HTRFlow, and several large language models (LLMs), including ChatGPT, Claude, and Gemini. Transkribus has shown limited success in accurately identifying handwritten placenames embedded within cartographic contexts. LLMs are being evaluated for their potential – already demonstrating impressive results; a testing environment is now under development to facilitate the systematic comparison of performance by different LLMs as well as integration with Transkribus and ground truth.

This work-in-progress explores feasibility for applying of AI towards complex historical documents. We invite feedback and collaboration from researchers with experience in machine learning, historical cartography, and archival digitization.

Specifically, we seek input on:

- Strategies for improving HTR performance in visually complex documents.
- Best practices for training and evaluating models on small, specialized corpora.
- Recommendations for performance metrics that capture both textual and symbolic recognition accuracy.
- Experiences with integrating LLMs - or other AI-driven tools - into workflows for historical document analysis.

By engaging with the community, we hope to refine our methodology and contribute to the development of more effective tools for the study of early modern cartography.

DT 2 — 08:30–10:00

08:30–09:00 **Hot Topics in the Parliament: A Topic Landscape Analysis with Emotion Expression**
Anna Ristilä

09:00–09:30 **A Community in Motion: Mapping Japanese American Migrations after World War II**
Saara Kekki

09:30–10:00 **Developing an Automatic Scansion Model for Early Modern Danish Verse**
Niels Nykrog, Manex Agirrezabal

08:30–09:00 *DEVELOPMENTAL TRACK* **[85]**

Hot Topics in the Parliament: A Topic Landscape Analysis with Emotion Expression

Anna Ristilä
University of Turku, Finland

Keywords: *topic landscape, concept, topic modeling, emotion analysis*

As text corpora grow beyond the limits of close reading, researchers increasingly rely on computational methods to uncover latent thematic structures and their social or linguistic correlates. Yet the terminology used to describe these analyses remains fragmented and method-specific. “Topic modeling”, “thematic categorization”, “semantic clustering” and similar labels refer to individual tools rather than to the broader analytical situation in which a corpus contains multiple topics whose relationships to metadata or contextual variables are of interest. In this paper, I introduce the concept “topic landscape” as a unifying and method-agnostic umbrella term for studies that seek to describe and interpret such relationships.

With topic landscape I refer to the overall distribution of topics within a corpus, augmented by one or more layers of variable correlation. Rather than focusing on the extraction methods of topics or variables (they may be readily encoded in the original data or derived in separate steps), the topic landscape perspective emphasizes the mapping of variation: how topics relate to contextual, temporal, or social dimensions. A topic landscape analysis may be either univariate, in which a single variable of interest (for example, speaker gender or party affiliation) is examined across all topics, or multivariate, in which multiple metadata dimensions jointly shape the landscape. This distinction encourages researchers to specify whether their focus lies on explaining a particular variable through topics (“topic landscape of a variable”) or on describing the corpus as a multidimensional thematic terrain (“topic landscape of the corpus”).

The term landscape is chosen deliberately. It conveys a spatial metaphor that helps conceptualize large-scale variation: topics form peaks and valleys whose colours and contours shift with variables

(time, speaker group, langauge etc.). Visual and statistical representations of such landscapes offer interpretable ways of approaching corpora that cannot be read exhaustively.

The term has been used before on papers doing research literature reviews with the help of topic modeling, but not in a very systematic way. For example, Lee, Seo and Geum (2018) use “topic landscape” in the title of their paper, which uses LDA to map topics in a corpus of research publications about product-service systems. However, in the rest of the paper they instead use the version “thematic landscape”. Use of “topic landscape” in similar settings is found on several papers, e.g. (Antons, Kleer and Salge 2015) and (Guo *et al.* 2018), but “thematic landscape” is used a bit more frequently, e.g. in (Thakral *et al.* 2023) and (Liu & McDonald 2019). On the other hand, not all papers with similar research settings mention either term.

While inspired by methods such as Structural Topic Modeling (Roberts et al. 2014), the topic landscape framework is explicitly method-agnostic. It does not prescribe any specific algorithms but provides a conceptual layer above them. By naming the research object – the multivariate distribution of topics – rather than the method, it helps bridge work conducted with diverse computational and statistical tools. This can foster better comparability and cross-disciplinary communication between studies in digital humanities, computational social science, and corpus linguistics.

By proposing the term topic landscape, this presentation seeks to open a discussion among researchers who study thematic variation in large text collections. The concept offers both a vocabulary and a perspective: a way to view text corpora not merely as sets of documents, but as terrains whose topography reveals how discourse, context, and social variables interact.

Building on this conceptual foundation, this paper presents an empirical use case that illustrates how the topic landscape analysis can be applied in practice. The demonstration focuses on Finnish parliamentary debates, a corpus characterized by both thematic richness and clear sociopolitical structure. The data includes plenary speeches from two decades, 2000-2020, previously analysed using Latent Dirichlet Allocation (LDA) to identify a set of recurring topics representing policy domains, social issues, and procedural matters such as "education", "taxation", or "development cooperation". The topic distribution is augmented with an emotion analysis (a more nuanced version of sentiment analysis; see e.g. Mao, Liu & Zhang 2024), constructing a topic landscape capturing how strongly different emotions are distributed across the topics.

Emotions expressed in political discourse offer an important dimension of variation – reflecting stance, framing, and rhetorical strategy – yet they are often studied separately from thematic structure. Here, an emotion model fine-tuned specifically for Finnish parliamentary speeches is applied, producing speech-level emotion scores across categories such as hopeful-optimistic-trust and fear-worry-distrust. The results demonstrate how a topic landscape analysis can reveal patterns that remain hidden when topics and emotions are studied independently. Certain policy areas – such as education and taxation – show distinctive emotional signatures, and certain historical events – such as the populist party vote landslide in 2011 – cause shifts in these signatures.

The integration of both topic and emotion layers enables a nuanced interpretation of parliamentary rhetoric: not merely what issues are discussed, but how they are emotionally constructed in discourse. The paper concludes by discussing methodological implications for future research. Through this use case, the presentation demonstrates the analytical potential of topic landscape thinking as a bridge between computational modeling and interpretive analysis.

In the developmental track, I would like to discuss the relevance, scope, and possible variations of definitions of 'topic landscape' from the point of view of using it in my forthcoming dissertation summary, and not just from the point of view of the paper in whcih I am currently using it. Is there anything problematic about the concept or its definition(s)? Does it overlap with something?

References

Antons, D., Kleer, R. & Salge, T.O. (2016). Mapping the Topic Landscape of JPIM, 1984–2013. J Prod Innov Manag, 33: 726-749. https://doi.org/10.1111/jpim.12300

Guo, L., Li, S., Lu, R., Yin, L., Gorson-Deruel, A. & King, L. (2018). The research topic landscape in the literature of social class and inequality. Plos One, July 2018. https://doi.org/10.1371/journal.pone.0199510

Hopp, C., Antons, D., Kaminski, J. & Salge, O. (2018). Perspective: The Topic Landscape of Disruption Research A Call for Consolidation, Reconciliation, and Generalization. Journal of Product Innovation Management, 35(3) 10.1111/jpim.12440.

Lee, H., Seo, H., & Geum, Y. (2018). Uncovering the Topic Landscape of Product-Service System Research: from Sustainability to Value Creation. Sustainability, 10(4), 911. https://doi.org/10.3390/su10040911

Liu, Y., Mai, F. & MacDonald, C. (2019). A Big-Data Approach to Understanding the Thematic Landscape of the Field of Business Ethics, 1982–2016. J Bus Ethics 160:127–150. https://doi.org/10.1007/s10551-018-3806-5

Mao, Y., Liu, Q., & Zhang, Y. (2024). Sentiment analysis methods, applications, and challenges: A systematic literature review. Journal of King Saud University - Computer and Information Sciences 36(4), 102048. https://doi.org/10.1016/j.jksuci.2024.102048

Roberts, M. E., Stewart, B. M., Tingley, D., Lucas, C., Leder-Luis, J., Gadarian, S. K., Albertson, B., & Rand, D. G. (2014). Structural Topic Models for Open-Ended Survey Responses. American Journal of Political Science, 58(4), 1064–1082. https://doi.org/10.1111/ajps.12103

Thakral, P., Srivastava, P.R., Dash, S.S., Jasimuddin, S.M., Zhang, Z. (2023). Trends in the thematic landscape of HR analytics research: a structural topic modeling approach. Management Decision, Vol. 61 No. 12 pp. 3665–3690. https://doi.org/10.1108/MD-01-2023-0080

09:00–09:30 *DEVELOPMENTAL TRACK* **[86]**

A Community in Motion: Mapping Japanese American Migrations after World War II

Saara Kekki

University of Helsinki, Finland

Keywords: *historical data, mapping, data matching*

This paper presents work in progress from *Community in Motion*, a Research Council of Finland–funded project (2023–2027) that investigates the post–World War II migrations of approximately 120,000 Japanese Americans following their incarceration and up to 1950. It is the first project to examine resettlement and subsequent mobility across the entire formerly incarcerated population rather than selected camps, families, or local communities. Up to now, the standard narrative has been that 50 percent of the former inmates dispersed to central and eastern parts of the United States, while another 50 percent returned to their former homes on the West Coast, where nearly all of them had lived before the war. However, my earlier research on a single concentration camp demonstrated that there was significantly more migration within the West Coast than previously acknowledged. Furthermore, return migration between immediate release and 1950 was significant, and I want to provide a more comprehensive and nuanced understanding of migration patterns and what drove people to particular communities, old or new.

The project is motivated by a historical question—how formerly incarcerated people rebuilt their lives spatially and socially in the immediate postwar period—but it is shaped by methodological and ethical concerns central to digital humanities scholarship. Specifically, it asks what becomes visible, distorted, or obscured when historians attempt to link and spatialize large-scale administrative datasets created under conditions of surveillance, displacement, and bureaucratic constraint.

The core of *Community in Motion* data consists of three datasets: the War Relocation Authority's 1942 camp entry records, the WRA's final departure rosters, and the 1950 U.S. census. The WRA datasets include structured individual and family identifiers and rich demographic information, covering the vast majority of the incarcerated population. The census, by contrast, was designed for an entirely different administrative purpose and lacks persistent identifiers, making it both indispensable and difficult to integrate. Since the 1950 census only became publicly available in 2022, its inclusion enables the first systematic examination of longer-term resettlement outcomes beyond the moment of camp departure.

This paper focuses on the process of matching individuals across these heterogeneous sources and treats record linkage as a site of historical interpretation rather than a purely technical preprocessing step. Working in collaboration with a data scientist, we employed two complementary approaches to person matching. The first is a deterministic, rule-based algorithm that identifies and ranks potential matches based on predefined conditions such as name similarity, birth year, and birthplace. The second is a supervised machine learning approach using gradient boosting, in which these and other features are weighted through training on a set of manually verified matches.

Rather than evaluating these approaches in terms of accuracy alone, the paper centers on error, uncertainty, and ambiguity. Thousands of cases remain unresolved or only partially matched, raising questions about confidence thresholds, false positives, and the consequences of exclusion. By examining concrete problem cases—such as name changes, inconsistent household composition, transcription errors, and culturally specific naming practices—I demonstrate how computational mismatch often reflects historically meaningful processes, including family separation, labor mobility, and ongoing surveillance of Japanese American communities after the war.

These methodological challenges directly inform the project's second major component: the development of interactive digital maps in collaboration with the Densho Digital Repository. The project aims to expand Densho's existing "Sites of Shame" platform beyond representative or typical routes to visualize individual and family migration trajectories at scale. Decisions about which matches to include, how to represent uncertainty, and how to avoid implying false precision are therefore both technical and ethical. Mapping, in this context, becomes not only a mode of dissemination but also an interpretive act that shapes historical narratives about return, belonging, and mobility.

In the presentation, I will also welcome a discussion on the ethical implications of large-scale historical data linkage and visualization. Although all of the sources used are publicly available, their combination produces more granular and invasive forms of knowledge than any single dataset alone. This raises questions about privacy, consent, and historical responsibility, particularly when working with communities that have experienced state surveillance and collective harm. I argue that ethical engagement in digital humanities requires attention not only to data access and anonymization, but also to how uncertainty, error, and incompleteness are communicated to users.

By foregrounding methodological failure and ethical deliberation, this talk positions computational tools as instruments for historical inquiry rather than solutions in themselves. It contributes to ongoing digital humanities discussions about responsible data reuse, interpretability, and the limits of automation, while offering a concrete case study of how historians can work productively with imperfect data to recover complex patterns of movement and resettlement in the aftermath of mass incarceration.

References

Friedman, Marissa, Cameron Ford, Mary Elings, Vijay Singh and Tracey Tan. (2021) Using AI/Machine Learning to Extract Data from Japanese American Confinement Records. UC Libraries Forum. https://doi.org/10.48448/c1rq-qf28.

Goeken, Ron, Lap Hyunh, T. A. Lynch, and Rebecca Vick. (2011) New Methods for Census Record Linking. Historical Methods vol. 44, 1: 7-14.

Kekki, Saara and Kai Ferragallo-Hawkins. (2025, in press) Using Digital Methods to Answer Humanities Questions about Migration. Muuttoliike - Migration vol. 51, 2.

Kekki, Saara. (2022) Japanese Americans at Heart Mountain: Networks, Power, and Everyday Life. Norman: University of Oklahoma Press. https://library.oapen.org/handle/20.500.12657/57770

Sites of Shame. maps.densho.org

09:30–10:00 *DEVELOPMENTAL TRACK* **[87]**

Developing an Automatic Scansion Model for Early Modern Danish Verse

Niels Nykrog[1,2], Manex Agirrezabal[1]

[1] *University of Copenhagen*

[2] *The Huygens Institute, Royal Netherlands Academy of Arts and Sciences*

Keywords: *Danish poetry, automatic scansion, metrics, verse history, early modern*

This paper discusses the challenges and opportunities of developing and using an automatic scansion model for analysing poetic rhythm in Danish verse literature from the period 15–1700. Our hypothesis is that such a model will enable a large-scale analysis of this seminal period of Danish literature, looking beyond the canon to explore an abundance of Danish verse texts written in genres such as religious hymns, ballads, occasional poetry, drama, and epics. This analysis would provide important new data about the rhythmic patterns of the period's literature, thereby enabling a pioneering statistical assessment of Scandinavian verse history, deploying a well-established method used fruitfully by scholars of phonologically similar languages. For a seminal introduction to the statistical study of verse,

see Gasparov and Tarlinskaja (1987). Notable recent applications of statistical methods and automated scansion models include Sisto (2023), Plecháč (2021), and Duffell (2008).

The project focuses on the period 15–1700 as an era of transition from free medieval verse forms ('knittelvers') towards the strictly regulated accentual-syllabic metrics that would come to define Danish poetry until the 20th century. This development of a new metrical standard in Danish has been described most often through canonical authors such as Anders Arrebo (1597–1637) and Thomas Kingo (1634–1703), leaving much to be said about its wider spread within what, following Margaret Cohen (1999, 22), we might term "the great unread" of early modern Danish literature

In order to build the scansion model, we are creating a small corpus (ca. 1,500 lines) of Danish verse, in which each line is annotated with a binary stress pattern. This will allow us to develop scansion tools following successful methodologies built for other similar languages, such as English (Agirrezabal et al., 2017) and Middle High German (Estes & Hench, 2016). Several scansion models have been created in the last few decades, for instance rule-based models that relied on lexical and morphosyntactic rules, Machine-Learning-based models, or Deep-Learning-based models. For a deeper insight into poetry analysis models, we refer the reader to the recent article by De Sisto et al. (2024). A highly successful architecture for building such models is the BiLSTM-CRF (Bidirectional Long Short-Term Memory Recurrent Neural Networks with Conditional Random Fields), because of its ability to model the relationships between letters (through internal letter representations in neural networks) and also between stresses (due to the Conditional Random Field layer). Once this model is built for Danish, we plan to perform a large-scale analysis of the poetic rhythm across a larger body of texts from 15–1700.

The project faces several challenges on which we would like feedback. First, the model we are planning to build will estimate the scansion of each verse line, where it implicitly (or jointly) performs syllable count and stress annotation. Syllable counting has been shown to be challenging in Large Language Models (Suvarna et al., 2024), as they have not been trained for that, and thus, it may also pose a challenge for our model. Second, the successful application of the model requires careful methodological consideration. As is often the case in digital literature studies, a significant challenge here is constructing a homogeneous textual corpus to analyse. While certain works from select Danish authors from the period 15–1700 have been digitized in various projects (e.g., *Arkiv for dansk litteratur* and *Tekstnet*), much work remains in order to construct a sufficiently broad textual corpus. Finally, we are interested in hearing any suggestions for other use cases for such a model outside large-scale text scansion and statistical analysis.

References

Agirrezabal, Manex, Iñaki Alegria, and Mans Hulden. "A Comparison of Feature-Based and Neural Scansion of Poetry." RANLP 2017 – Recent Advances in Natural Language Processing Meet Deep Learning, Incoma Ltd., Shoumen, Bulgaria, November 10, 2017, pp. 18–23. https://doi.org/10.26615/978-954-452-049-6_003.

Cohen, Margaret. The Sentimental Education of the Novel. Princeton University Press, 1999.

De Sisto, Mirella, Hernández-Lorenzo, L., De la Rosa, J., Ros, S., & González-Blanco, E. "Understanding Poetry Using Natural Language Processing Tools: A Survey." Digital Scholarship in the Humanities, vol. 39, no. 2, 2024, pp. 500–521.

De Sisto, Mirella. "The Development of a Poetic Tradition. A Study of a Dutch Renaissance Poetry Corpus." Studia Metrica et Poetica, vol. 10, no. 1, 2023, pp. 36–68.

Duffell, Martin. A New History of English Metre. Routledge, 2008.

Estes, Alex, and Christopher Hench. "Supervised Machine Learning for Hybrid Meter." Proceedings of the Fifth Workshop on Computational Linguistics for Literature, Association for Computational Linguistics, 2016, pp. 1–8. https://doi.org/10.18653/v1/W16-0201.

Gasparov, Mikhail Leonovich, and Marina Tarlinskaja. "A Probability Model of Verse (English, Latin, French, Italian, Spanish, Portuguese)." New Metrics, vol. 21, no. 3, 1987, pp. 322–58.

Plecháč, Petr. Versification and Authorship Attribution. Prague: Karolinum Press, 2021.

Suvarna, Ashima, Harshita Khandelwal, and Nanyun Peng. "PhonologyBench: Evaluating Phonological Skills of Large Language Models." Proceedings of the 1st Workshop on Towards Knowledgeable Language Models (KnowLLM 2024), Association for Computational Linguistics, 2024, pp. 1–14. https://doi.org/10.18653/v1/2024.knowllm-1.1.

Session 5A — 08:30–10:30

08:30–09:00 *LONG PAPER* **[88]**

Retouched but Not Restored: Using a Generative Model to Enhance Noisy Newspaper OCR

Lina Samuelsson[1], Daniel Brodén[2], David Alfter[2], Johan Malmstedt[2]
[1] *Mälardalens universitet, Sweden*
[2] *Göteborgs universitet, Sweden*

Keywords: *generative models, OCR, digital epistemology*

Digitized historical newspapers remain indispensable sources for humanities research (Gooding 2018), yet their value is often limited by OCR noise, which hampers close reading and large-scale analysis (cf. Löfgren and Dannélls 2024). This presentation reports an exploratory study applying a generative language model to "retouch" OCR-degraded Swedish book reviews from 1905–6 in the National Library of Sweden's (Kungliga biblioteket; hereafter KB) newspaper collections. At issue is not only the technical problem of noise but also the epistemic question of mediation: how generative systems reshape what counts as source material in digital textuality. As Ingvarsson (2020) notes, digitisation reconfigures our modes of knowing as much as our modes of seeing. Our experiment thus positions itself at the intersection between enhancement and epistemic displacement.

Our aim is twofold: (i) to examine to what extent zero-shot prompting can make noisy OCR more usable for natural language processing (NLP) purposes; and (ii) to reflect critically on the epistemological consequences of deploying probabilistic, largely opaque systems for this purpose.

The case study is part of the mixed-methods research project The Order of Criticism Revisited (Ingvarsson et al. 2022), which investigates the interplay between "traditional" literary scholarship and computational methods on both practical and epistemological levels (see e.g., Brodén et al. 2024). In the project, we collected approximately 5,800 reviews from Swedish newspapers, but suboptimal digitization in KB's newspaper collections has led to substantial OCR noise and segmentation errors (Jarlbrink et al. 2016; Börjesson et al. 2024; Sikora and Haffenden 2024). For copyright reasons, we narrowed our focus to reviews from 1905–6 and to a smaller subset for which manual transcriptions are available (see Samuelsson 2013).

Rather than fine-tuning a model, for this case study, we used zero-shot prompting to instruct GPT-4o to act as an expert in historical OCR correction, cleaning errors while preserving historical spelling and style (see also Brodén et al., in press). For evaluation, we compared (i) GPT-4o's output, (ii) KB's raw OCR of the newspaper text, and (iii) manual "gold" transcriptions of the original text in the digitized facsimiles. We applied the pipeline to the full set of reviews (n=394) and to the manually transcribed subset (n=21).

We frame our discussion within Espositio's characterization of AI as "artificial communication" rather than "artificial intelligence" (Espositio 2022), treating GPT-4o's outputs as communicatively calibrated responses rather than intelligence-based reconstructions (Fazi 2019). We also draw on Romanyshyn's

(1989) notion of distant technology to emphasise how these generative models may extend technological possibilities while introducing methodological distance through their probabilistic character and limited transparency. While the model's output may appear more "correct", it is shaped by computational logics that are generically inferred and largely inaccessible. Bringing these perspectives together highlights the model's ambiguous role as a tool and underscores the need for methodological and epistemological reflection.

We argue that the model's OCR output is better understood as "retouching" rather than "correction," (cf. Boros et al. 2024) "reconstruction," or "restoration," as the generated text derives not from the original print but from a flawed digitized version. While the unprocessed OCR version retains an indexical link to the original newspaper page, traceable to the physical imprint of ink on paper, the GPT-retouched version lacks this causal – and, in our context, reliable – connection (cf. Sterne 2016). The result is a statistically inferred approximation that resembles the original but is generated without reference to the original, mediated instead by an algorithm oriented toward probabilistic pattern recognition rather than reproduction.

For a closer comparison of how much the retouched reviews differ from the original text in the newspapers, we have compared their Levenshtein distance between the reviews as a simple, interpretable measure of divergence.

Across the full set, the OCR-retouch transformation required on average ~320 character-level operations per text (95% CI: 253–386; n=394), indicating substantial change relative to raw OCR. On the smaller, transcribed subset, GPT-4o retouchings were, on average, ~104 operations closer to the gold standard than the raw OCR (95% CI: 64–143; n=21). In sum, the generated outputs typically approximate the manually transcribed texts more closely than the unretouched OCR does, suggesting meaningful gains in legibility and downstream analyzability for tasks such as search, concordance, and topic discovery.

However, while retouching thus performs well overall, it is not exact and we therefore also need to take into account how it differs from the original. A comparison between each and one of the individual manually transcribed texts and GPT-4o's output shows that the differences range from as few as eight characters to a maximum of 446. In the latter case, however, this is in a text totaling 8,746 characters, which thus amounts to no more than 5% of the text.

Notably, larger alterations in the texts, such as when several words or entire sentences have been changed, seem to occur when the print quality in the newspapers is particularly poor and the OCR particularly noisy. It is at times like these that the model takes great liberties in filling out the text. On the other hand, the raw OCR in these cases would be virtually unreadable.

Thus, the retouching works well from a quantitative perspective and makes them more readable and useful for computation. However, since it is still a matter of retouching, and the changes are made based on probabilistic calculations, we need to take a closer look at what kinds of changes the model makes and what significance these may have for the continued analytical work.

A sentence-by-sentence comparison of the retouched reviews with the original texts shows that the differences are overall limited. A common difference is that the generative model modernizes spelling. While these alterations do not affect comprehension, they constitute quantitative differences between the texts. From a NLP perspective, this could also become problematic if different spelling variants are not taken into account.

Of lesser importance are also differences in punctuation, which are common. Commas are placed differently, hyphens are used by GPT where the original lacks them, and quotation marks end up in the wrong places. While these issues usually do not affect the overall understanding of this type of material, it obviously matters.

Since the retouching is based on probabilistic estimation, many of the errors involve words being replaced with terms that are essentially synonymous (e.g., "gnällande" (whining) becomes "grinande" (crying) or being given a different preposition while still conveying nearly the same meaning (e.g., "tomt om" (empty of) becoming "tomt för" (empty for)). In other cases, the model seems to misread characters, resulting in words that don't even exist (e.g., "grus" (gravel) becomes "grua").

While these types of errors are usually easy for a human reader to notice and check against the original, more insidious cases are where the misread words become actual but incorrect words (e.g., "gatorna"

(the streets) becoming “gåtorna” (the riddles), and “resa” (travel) becoming ”rasa” (collapse)). Sometimes the context makes it clear to a human reader that the word must be wrong (e.g., “örn” (eagle) instead of “om” (about)), but in other cases, it may mislead the reader and is obviously also problematic in NLP analyses based on word frequencies as well as for more contextual models, if such errors occur on a large scale.

As an example of places where the model has filled in significantly longer sections, these often also do not have crucial importance; one example of this is “[s]å kommer den arbetsamma utgången och epilogen” (“[t]hus comes the laborious conclusion and the epilogue”) that has become “[s]å kommer den arbetet afslutande epilogen” (“[t]hus comes the work-concluding epilogue”), while the model's choice of the term “människofienden” (“enemy of mankind”) being changed to “mannen med kniven” (“the man with the knife”) represents a significantly greater shift in meaning, though not particularly crucial for an interpretation of the review as such.

We conclude by synthesizing our findings: although GPT-4o can produce what appears to be high-quality OCR cleaning, its outputs are best regarded as retouchings rather than restorations or reconstructions. Generative retouching thus performs a second abstraction. As Eve (2024) observes, OCR already reduces the semiotic richness of the printed page by translating visual inscription into symbolic text. The generative model extends this reduction, operating not on the level of letters but through probabilistic relations of meaning. Rather than restoring what OCR has lost, it adds another layer of mediation, transforming one abstraction into another. The result is a twice-filtered representation: more legible and analytically tractable, yet increasingly detached from the physical artifact. What appears as textual improvement, coincides with a loss of indexical depth, as the material traces of print give way to a statistically inferred approximation.

We also draw together our theoretical argument, contending that generative models remain methodologically distant and therefore require both continued refinement of practical applications and a deepened epistemological consideration of the conditions under which these tools operate. Bringing these perspectives together, we suggest considering GPT-4o’s intervention less as a corrective act and more as a communicative simulation that reshapes evidential relations. In this sense, the generative model becomes both an analytic aid and an epistemological filter.

References

Boros, Emanuela, Maud Ehrmann, Matteo Romanello, Sven Najem-Meyer, and Frédéric Kaplan. 2024. “Post-Correction of Historical Text Transcripts with Large Language Models: An Exploratory Study.” In Proceedings of the 8th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature (LaTeCH-CLfL 2024. Association for Computational Linguistics: 133–159.

Brodén, Daniel, Jonas Ingvarsson, Lina Samuelsson, and Victor Wåhlstrand Skärström. 2024. “Visualization as Defamiliarization: Mixed-Methods Approaches to Historical Book Reviews.” Journal of Computational Literary Studies 3 (1): 1–26.

Brodén, Daniel, Lina Samuelsson, and David Alfter. In press. “Retouching and Refiguring Literary Criticism: Experiments with a Generative Model for Analyzing Book Reviews”, Flows & Frictions: Mixed Methods for AI-Driven Research on Historical Media, edited by Daniel Brodén, and Lina Samuelsson, LIR.skrifter

Börjesson, Love, Chris Haffenden, Martin Malmsten, Fredrik Klingwall, Emma Rende, Robin Kurtz, and Faton Rekathati et al. 2024. “Transfiguring the Library as Digital Research Infrastructure: Making KBLab at the National Library of Sweden.” College & Research Libraries 85 (4): 564–82.

Esposito, Elena. 2022. Artificial Communication: How Algorithms Produce Social Intelligence. MIT Press.

Eve, Martin Paul. 2024. Theses on the Metaphors of Digital-Textual History. Stanford, CA: Stanford University Press.

Fazi, Beatrice. 2019. “Can a Machine Think (Anything New)? Automation Beyond Simulation.” AI & Society: Knowledge, Culture and Communication 34 (4): 813–824.

Gooding, Paul. 2019. Historic newspapers in the digital age: “Search all about it”. Routledge.

Ingvarsson, Jonas. 2020. Towards a Digital Epistemology: Aesthetics and Modes of Thought in Early Modernity and the Present Age. Palgrave Macmillan.

Ingvarsson, Jonas, Daniel Brodén, Lina Samuelsson, Victor Wåhlstrand Skärström, and Niklas Zechner. 2022. “The New Order of Criticism: Explorations of Book Reviews Between the Interpretative and Algorithmic.” In The 6th Digital Humanities in the Nordic and Baltic Countries Conference (DHNB 2022), edited by Karl Berglund, Matti La Mela, and Inge Zwart. CEUR-WS.org: 228–34.

Jarlbrink, Johan, Pelle Snickars, and Cristian Colliander. 2016. “Maskinläsning: Om massdigitalisering, digitala metoder och svensk dagspress.” Nordicom Information 38 (3): 27–40.

Löfgren, Viktoria, and Dana Dannélls. 2024. “Post-OCR Correction of Digitized Swedish Newspapers with ByT5.” In Proceedings of the 8th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature (LaTeCH-CLfL 2024), edited by Juri Bizzoni, Stefania Degeatano-Ortlieb, Anna Kazantseva, and Stan Szpakowicz. Association for Computational Linguistics: 237–42.

Romanyshyn, Robert. 1989. Technology as Symptom and Dream. Routledge.

Samuelsson, Lina. 2013. Kritikens ordning: Svenska bokrecensioner 1906, 1956, 2006. Bild, text & form.

Sikora, Justyna, and Chris Haffenden. 2024. “AI, Data Curation and the Readiness of Heritage Data: Exploring the Swedish Newspaper Archive at KBLab.” In Proceedings of the Huminfra Conference (HiC 2024), edited by Elena Volodina, Gerlof Bouma, Markus Forsberg, Dimitrios Kokkinakis, David Alfter, Mats Fridlund, and Christian Horn et al. Linköping Electronic Conference Proceedings: 60–66.

Sterne, Jonathan. 2016. “Analog.” In Digital Keywords: A Vocabulary of Information Society and Culture, edited by Benjamin Peters. Princeton University Press: 31–44.

09:00–09:20 *SHORT PAPER* **[89]**

Printing Possibility: Scaling inconspicuous labour advertisements to reconstruct early modern Danish labour markets, 1750-1850

Sofus Landor Dam

Aalborg University, Denmark

Keywords: *Early modern labour markets, digital reconstruction, gender, long-term structures, geographic difference*

During late Danish absolutism, buying and selling labour power in advertisement newspapers became commonplace, an everyday phenomenon that slowly grew in size with the popularity of print periodicals. Each labour advertisement follows a stable format that contains information on a range of aspects: the purpose of the advertisement (reformulated as supply and demand), the position in question (and, by proxy, gender), experience with earlier occupations, qualifications, and requirements for testimonials from the previous employer (so-called Skudsmål). Reading the advertisements draws attention to the coercive structures engrained in the contemporary labour market. Unemployment and vagrancy was illegal, and securing employment was contingent on exiting previous service lawfully (i.e. sanctioned by your master) and with good Skudsmål on hand, to attest your proper work ethic and fidelity. The existence of these themes are not a new discovery, however, but have been studied primarily qualitatively or at lower scales in the past (see for example Østhus 2013 and Friis 1964). Consequently, at the individual level, each labour advertisement rarely tells us something new or surprising. The sheer ubiquity and stability of the genre instead allows another use for the material. When scaled up with enough individual advertisements, we can empirically investigate long-term structures in early modern labour markets across variables such as time, geography, occupation and gender. This level of scale is within reach due to a newly digitized text corpus consisting of approximately 4.9 million newspaper texts, covering the largest Danish and Norwegian newspapers from before 1900 (see Heinsen and Bøgeskov 2025). Allying with Johan Heinsen, we trained a binary classifier model to identify the existing labour advertisements in the material (trained on a 1000-row manually tagged random sample from the entire newspaper corpus). This model found c. 241.000 advertisements (accuracy: 0.99, F1: 0.94. See Heinsen 2025a). The initial problem was that labour advertisements often appear alongside each other under headers such as "seeking service", or nested within other types of texts in heterogenous "announcement"-rubrices in the newspapers. The text segmentation of the original dataset frequently failed to correctly split these texts, owing largely to the fact that the segmentation algorithm was not tailored for the labour advertisements.This meant two things: 1) the positively identified ads contained a large amount of non-relevant text, and 2) the true amount of labour ads was likely much higher than the initial classification suggested. To combat these two challenges, I created a bespoke tokenizer fitted to the most frequent and common lingual indicators of beginnings and endings of advertisements, splitting the texts into smaller bites resembling sentences. Afterwards, I ran our classifier model again, and filtered off the text bits that were not positively identified as labour advertisements. This data structure far more closely resembles a "one row, one instance"-dataset, with an estimated success rate of 96.12% (based on a manually validated random sample, where success meant that a given text contained only 1 labour advertisement, even if fragmented; n = 500). This data transformation resulted in a dataset with 344.000 labour advertisements, which we further enriched using binary classifiers to spot gender and supply/demand status (see Heinsen 2025b and Heinsen and Dam 2025 for model

metrics). Corpus-wide counts show an overweight of women across the entire period (64% versus 36% men), primarily concentrated in Copenhagen newspapers as labour supply, whereas provincial newspapers show distributions that predominantly favor a demand for men. The influx of supply labour advertisements in the Copenhagen started around the last decade of the 1700's, continuing into the 1800's and until the end of the period in 1850, even while other Danish regions experienced falling amounts of both demand and supply labour advertisements, for example during the economic crisis in the aftermath of the Napoleonic Wars. Thus, there are clear signs of highly different types of labour markets across the Danish realm, where the urban metropolitan center of Copenhagen reads as a market with relatively high turnover, perhaps indicating higher degrees of labour precarity and scarcity, especially for women, compared to provincial towns. This is also reflected in the fact that the general supply was almost exclusively higher than demand throughout the period. That is likely partly due to Københavns Adresseavis being oriented towards facilitating the contact between employer and worker, thereby cementing its centraility in the metropolitan labour market. However, when compared to census data in the year 1837, the amount of labour ads outweigh the amount of persons registered in positions mentioned in the ads (~10.500 vs. ~12.500 ads). This can either mean enormous turnover rates or increased mobility; regardless, it testifies to an extemely active labour market. In regards to mobility, many advertisements explicitly mentions either the origins of a worker or the desired origins of a potential employee. There is considerable weight on people from the countryside, "the Provinces", as well as Jutland and the duchy of Holsten. The focus on rural origins in both supply and demand advertisements highlights that mobility towards towns and the urbanity of Copenhagen could likely have given workers an edge in these labour markets. This is in line with findings of Løgstrup (1987) Henningsen (1995), who has documented that estate owners from Jutland sometimes feared that their escaped servants would go to Copenhagen (or Holsten due to better pay), and that their skills with threshing was sought after. In conclusion, the computational and large-scale examination of labour advertisements both confirms and nuances prior research results, as well as opening ways of empirically researching specific questions related to gender, mobility, regionality and supply and demand.

References

Heinsen, J., and Bøgeskov. C. 2025. A World in Print: Introducing a Danish-Norwegian corpus of historical newspapers [in English]. Working paper. arXiv. https://doi.org/10.48550/arXiv.2509.02356.

Heinsen, J. 2025a. "Labour advertisement identifier". URL: https://huggingface.co/JohanHeinsen/Labour_advertisement_identifier. HuggingFace.

Heinsen, J. 2025b. "Labour ads demand". URL: https://huggingface.co/JohanHeinsen/Labour_ads_demand.

HuggingFace.

Heinsen, J. & Dam, S. L. 2025. "Labour ads gender". URL: https://huggingface.co/JohanHeinsen/Labour_ads_gender

HuggingFace.

Østhus, H.. 2013. “Contested Authority: Master and Servant in Copenhagen and Christiania, 1750-1850.” PhD diss., European University Institute.

Henningsen, P. 1995. "Hedens Hemmeligheder: Livsvilkår i Vestjylland 1750-1900". Overgaard Bøger.

Friis, L. 1964. "Skudsmålsbøgerne som kilde til kvindelige tjenestetyendes forhold". Unpublished master's thesis. University of Copenhagen.

Løgstrup, B. 1987. "Bundet til jorden: Stavnsbåndet i praksis 1733-1788". Landbohistorisk Selskab.

09:20–09:50 *LONG PAPER* **[90]**

“Were He Not Called Sigurðr” – Personal Name Case Studies from Medieval Norway

Elisabeth Maria Magin

University of Oslo, Norway

Keywords: *runes, runology, onomastics, Old Norse, medieval studies*

Sigurd the Dragonslayer, known as Siegfried in the original Middle High German Nibelungenlied and Sigurðr in the Old Norse sources, clearly left an impression on medieval Norwegians; enough so that the name can be found no less than 19 times in the corpus of almost 700 runic inscriptions from Bergen,

Norway. Only one other name appears equally frequently: María. There is, however, a great difference between how these two names are used: Sigurðr appears exclusively as a personal name, carried by someone living in or visiting Bergen. Conversely, analysis of the inscription content shows that María exclusively refers to the Virgin Mary. Other than Sigurðr, in medieval Bergen María was a taboo name – a saint that can be addressed, but not a name that may be given to a child.

This paper presents the data model and results of a SQL database-supported study of the almost 250 personal names found in the Bergen runic inscriptions dating to between ca. 1120 to 1450. It critically considers their cultural, archaeological and textual context and uses case studies to demonstrate the varied contexts in which personal names appear in runic inscriptions, ranging from invocations of saints to boastful declarations of sexual experiences. Using the fire layer chronology and GIS, the paper also discusses possible changes in naming customs over the ca. 300 years from the founding of the town until the late 15th century.

Within the field of name studies in Norway, medieval personal names, naming customs and name use have received little attention, with some of the most relevant overviews (Halvorsen 1981; Halvorsen 1984; Johannessen 2002) now being several decades old. This may partly be owed to the fact that the most comprehensive collection of medieval names from written sources, the name lexicon *Norsk-Isländska Dopnamn Ock Fingerade Namn* (Lind 1905-1915; Lind 1931) in itself is over 100 years old by this point, and has never been properly digitised. Attempts at using the lexicon to infer family background and place of origin for the Bergen runecarvers (Hagland 1988a; Hagland 1988b; Hagland 1989) have thus also been met with scepticism and critique (Seim 1989; Seim 1991).

In part, this is owed to the lack of written sources for the immediate post-Viking Age period in Norway. Sufficiently large numbers of manuscript sources are lacking up until the 1300s, and the names occurring in the Kings' and Icelanders' sagas, which describe events during the Viking Age (ca. 800-1100) in Scandinavia, have long been regarded with scepticism due to the 300-year gap between the setting of the saga and the manuscript being written down. The runic inscriptions, which have been dated to the time period between approximately 1120 and 1450, are, however, genuine and primary sources of name occurrence and usage in medieval Bergen. They can thus provide insight into personal names and naming for a point in time not just when manuscript sources are scarce, but also serve as a comparative corpus for when manuscript writing has become established in Norway post ca. 1200. As the administrative centre and largest trading town in Norway during the medieval period, the runic material from Bergen is also uniquely suited to provide insights into phenomena such as visiting tradespeople from other countries (Germany, the Netherlands) and the use of writing in a two-script society (Latin and Runic), where especially members of the upper echelons of society might be choosing their medium of writing depending on circumstances.

Methodologically, it is difficult to determine not just the frequency of a specific name within any given corpus of names, but also its use within a specific level of society. The number of inscriptions (ca. 700) and the total count of personal names (> 400), not to mention the necessity to consider the archaeological context, additionally complicate analyses of the material. In order to facilitate analyses, a complex SQL-based data model was developed (Magin 2023). Amongst other aspects of the runic inscriptions, it allows storing information on:

- the inscription content and context the name appears in;
- whether the name, based on this, is considered to refer to an actual person or an invocation of a saint/deity/supernatural being;
- whether it appears as a byname, patronym or metronym;
- if possible to determine, which language the name appears in ("Siegfried" would be Middle High German, whereas "Sigurðr" is Old Norse and Sigurd modern Norwegian);
- the frequency of each name in other written sources, predominantly the diplomas from Iceland and Norway (*Diplomatarium norvegicum* 1848-1920; *Diplomatarium islandicum* 1857-1952).

The paper firstly aims to provide a short overview over the current state of scholarship concerning names and naming customs in Norway from the Viking Age to the 15th century. Particular attention will be paid to the retention of Old Norse names with close ties to pre-Christian religion and myth in Norway (such as Sigurðr), and the introduction of Christianity-inspired names (such as María) into the name repertoire post-Conversion.

In the second section, the paper discusses how the onomastic part of the database has been modelled to facilitate storing the above-mentioned information about each single name, and how it connects to the archaeological database, where information on the find circumstances, dating and the object itself are stored. Due to frequent fires visible as ash layers in the excavation, the dating of objects tends to be quite precise, in several cases within two decades. It thus becomes possible to group the names into phases between fire layers, and although uncertainties must be considered (for example, objects can have been in use used prior to the fire already, but were only deposited a few years after, when they broke or became otherwise superfluous), the fire layers allows for an unusually detailed chronology of name frequency and use in medieval Bergen.

In the last section, some of these observable trends are discussed more generally, such as the question of when Christianity-inspired names start appearing in the runic material from Bergen, not just as invocations of saints, but as genuine personal names which are now in active use within the town's population. Two case studies are presented to illustrate the developments, and finally, the potential sociological implications of these changes are discussed as well – what could have sparked these trends? How might naming customs have changed, and why?

As a final note, the paper briefly discusses the question of how reliable any conclusions based on a comparison of names from epigraphic and manuscript sources are, especially when their timeframes only overlap partially. However, despite remaining doubts and uncertainties concerning the material, the paper hopes to demonstrate that, by using a flexible, SQL-based data model, comparisons on a broader basis become not only more manageable, but can significantly contribute to our understanding of developments in personal naming even for time periods when the source material is scarce.

References

Diplomatarium islandicum: Íslenzkt fornbréfasafn, sem hefir inni að halda bréf og gjörninga, dóma og máldaga, og aðrar skrár, er snerta ísland eða íslenzka menn. 1857-1952. Íslenzkt Fornbréfasafn. Reykjavík: Hið Íslenska bókmenntafélag.

Diplomatarium norvegicum. 1848-1920. Oslo: Kommisjonen for Diplomatarium Norvegicum.

Hagland, Jan Ragnar. 1988a. Nokre onomastiske sider ved runematerialet frå bygrunnen i Trondheim og Bryggen i Bergen. Studia Anthroponymica Scandinavica: 13–25.

Hagland, Jan Ragnar. 1988b. Runematerialet frå gravingane i Trondheim og Bergen som kjelder til islandshandelens historie i mellomalderen. Historisk Tidsskrift: 145–156.

Hagland, Jan Ragnar. 1989. Islands eldste runetradisjon i lys av nye funn frå Trondheim og Bergen. Arkiv för Nordisk Filologi: 89–102.

Halvorsen, Eyvind Fjeld. 1981. Personnavn. Island og Norge. In Kulturhistorisk leksikon for nordisk middelalder, 199–206. Kulturhistorisk Leksikon for Nordisk Middelalder 13.

Halvorsen, Eyvind Fjeld. 1984. Innlån av fremmede personnavn i Norge i tidlig gammelnorsk tid. In Festskrift til Ludvig Holm-Olsen på hans 70-årsdag den 9. Jun 1984, ed. Bjarne Fidjestøl, 114–123. Øvre Ervik: Alvheim & Eide.

Johannessen, Ole-Jørgen. 2002. Kristne personnavn i norsk middelalder. In I: Kristendommens indflydelse på nordisk navngivning. Rapport fra NORNAs 28. Symposium i skålholt 25.–28. Maj 2000, ed. Svavar Sigmundsson, 29–57. NORNA-Rapporter 74. Uppsala: NORNA-förlaget.

Lind, Erik Henrik. 1905-1915. Norsk-isländska dopnamn ock fingerade namn från medeltiden. Uppsala: Lundequistska bokhandeln.

Lind, Erik Henrik. 1931. Norsk-isländska namn ock fingerade namn från medeltiden. supplementband. Uppsala: Lundequistska bokhandeln.

Magin, Elisabeth Maria. 2023. Data-based runes. Macrostudies on the Bryggen runic inscriptions. The Bryggen Papers 10. The University Museum of Bergen & The Faculty of Humanities, University of Bergen. https://doi.org/10.15845/bryggen.v100.

Seim, Karin Fjellhammer. 1989. Runeinnskrifter fra Trondheim og Bergen som kilder til Islandshandelens historie? Et innfløkt proveniens-spørsmål. Historisk Tidsskrift: 333–347.

Seim, Karin Fjellhammer. 1991. Flere onomastiske sider ved runematerialet fra bygrunnen i Trondheim og Bryggen i Bergen. Studia Anthroponymica Scandinavica: 21–32.

09:50–10:10 *SHORT PAPER* **[91]**

Named Entity Recognition in the Historical Meeting Protocols of the Tartu City Council

Siim Orasmaa, Kadri Muischnek, Sofia Kriuchkova
University of Tartu, Estonia

Keywords: *historical language processing, corpus annotation, named entity recognition*

In this paper, we describe our efforts to enable automatic named entity recognition in the historical meeting protocols of the Tartu City Council from 1918–1940, written in Estonian. We describe the conversion of automatically transcribed protocol pages to a form suitable for manual corpus annotation and model building, introduce our choice of name categories, and describe the manual annotation process. The resulting corpus is annotated for 8 named entity categories, with fairly good inter-annotator agreement on the majority of the categories. Finally, we present experiments on fine-tuning two BERT-like models on the dataset, and report that the best model achieved an overall F1-score of 0.8246.

10:10–10:30 *SHORT PAPER* **[92]**

Women in Business on the Rhine in the 18th Century: Comparison of Business Patterns and Quality Control of Identification after AI-Supported Named Entity Recognition (NER)

Lauri Matias Heinonen
University of Bamberg, Germany

Keywords: *Named entity recognition (NER), customs register, semi-structured data, business history, early modern period, female entrepreneurship, female work, 18th century, river trade, inland trade, Rhine trade, tariffs, Germany, Netherlands*

This short paper of 13 pages for the DHNB 2026 conference studies how female shipmasters and other professionals can be identified in the customs registers of Schenkenschans customs station (1630-1810) in the 18th century at the border between the Netherlands and Germany through Named Entity Recognition (NER). The paper uses a study of widows to review methodologies of NER on an early modern dataset, the customs registers of Schenkenschans (1630-1810), a historical fortress and customs station on the Rhine on the border between Netherlands and what today is Germany.

I conduct Named Entity Recognition (NER) on these data through two different methods for comparison, Levenshtein Distance and Ratcliff/Obershelp string matching algorithm, to identify the female widows who are the female entrepreneurs in the data. These two methods are then applied to matching the widows with their deceased husbands. In this step, I compare the selection of products widows, and their deceased husbands transported through the Schenkenschans and by the locations from where their ships came from. This shows continuities and discontinuities in trade patterns between the widows and their deceased husbands.

The identification of the widows is based on the way they were registered in the dataset: widows in my Dutch-language material had the title weedwe (widow) or some of its spelling variants as their title followed by the name of their deceased husband. Thus, I apply the methods of Levenshtein Distance and Rattcliff/Obershelp string matching algorithm on the RoleName and PersName tags of the source XML files. Here is an example on the RoleName tag and the PersName tag and how they work in the example line 367 in the XML file for the customs register of 1737-1738:

<roleName>De Wed:

Verkerck</persName></roleName>

The string-matching algorithms use the corpus defined by the author to find the RoleName tags including the title weedwe or some of its spelling variations, in this case Wed:. After this, they match the title they find with the person's name in the PersName tag that follows as this is the last name or the full name of the deceased husband. In this case, the name of the widow would be identified by the algorithms as "Verkerck".

In the early modern period, women were allowed to run a business if their husband had run a business and passed the business on their widow after their death. See Erickson (2024) on the 18th century UK and Moring (2025) on the 19th century Finland. Earlier literature points out that there are clear research gaps in business history literature. Wilson et al. (2022) find major challenges to the field of business

history despite flourishing research that include, for example, a lack of overarching paradigms in the field, limited teaching compared to research and neglecting some research themes such as gender questions.

Other research like Gelderblom and Trivellato (2018) also find that literature has neglected women and their economic contribution especially in literature focusing on the early modern period. The temporal focus on only some periods of history comes up in other contributions as well. Banken and Ressel (2024) discuss the state of German business history as a research field and find that research into German business history has neglected the period of early industrialization between 1850 and 1870 and the whole period prior to 1815.

The empirical results of comparing the business patterns of the widows and their deceased husbands are as follows:

The analysis of the female, widowed entrepreneurs and their deceased husbands shows some main continuities and discontinuities between them. Firstly, the female widows seemed to have both continued and discontinued the businesses of their deceased husbands in the same geographical places in Germany and the Netherlands. This logical given that the source data used in this paper are the customs registers of a single customs station, meaning that a relatively low level of diversity and change in geographical locations of trade is to be expected.

There are, nevertheless, some geographical patterns. It seems that Cologne became more important for female shipmasters increased in the late 18th century and Amsterdam less important. Cologne was an important city for the widows both as a new, a continued and a discontinued location after the death of the husband.

An issue in the current research design of this paper is that the direction of ships going past the Schenkenschans customs station in 1737-1810 is now not included in the analysis of locations. While this data is available, the analysis here has been made simpler due to the limitations on the length of the paper.

When analyzing the continuities and discontinuities in the selection of products that the widows and their husbands, it seems there are more discontinuities than with the geographical locations. It seems that the trade in tobacco and associated products like the pijpaarde soil conducted by the female, widowed entrepreneurs started to increase on the Rhine from the 1750s onwards.

When widows started to trade new products that their deceased husbands had not traded, it seems wine was an important new product in their sortiment until the 1760s. However, from the 1770s onwards, widows started to conduct more trade with coffee on their own. Additionally, the widows started to conduct new trade in construction and raw materials like stone and iron after their husbands had died.

The discontinuities by the widows after their husband’s death relate to two main categories of products. Firstly, widows tended to discontinue very capital-intensive trade like timber. Additionally, it was common for them to discontinue trade in very common products like pottery and dried goods. This reflects the extensive trade with these products in general, meaning there were many entrepreneurs both starting and quitting the business with such products.

It is worth noting, however, that the changes in products sold by the widows might not only reflect their own business decisions. They might reflect larger structural changes of the Rhine economy. The approach used here does not allow assessing the cause of the changes.

References

Ågren, Maria, Sofia Ling, and Linnea Henningsson, 'Marriage and Work: Continuity and Subtle Change', in Maria Ågren (ed.), Gender, Work, and the Transition to Modernity in Northwestern Europe, 1720–1880 (Oxford, 2025; online edn, Oxford Academic, 22 Jan. 2025), https://doi.org/10.1093/9780198934325.003.0005, accessed 4 Feb. 2026.

Banken, Ralf and Ressel, Magnus (2024). “Unternehmensgeschichte ist mehr als Unternehmenszeitgeschichte - Plädoyer für die Wiederentdeckung des 18. und frühen 19. Jahrhunderts” (Business history is more than contemporary business history – Appeal for rediscovery of the 18th and early 19th centuries), pp. 83-104. In Kleinöder, Nina, Marx, Christian, Gehlen, Boris and Czierpka, Juliane (eds). Neue Perspektiven der Unternehmensgeschichte (New perspectives of business history). Brill, Schöningh, pp. 1-350. DOI: https://doi.org/10.30965/9783657794775.

Erickson, Ama Louise (2024). Wealthy businesswomen, marriage and succession in eighteenth-century London. Business History, 66(1), 29–58. https://doi.org/10.1080/00076791.2022.2036131

Gelderblom, Oscar, and Trivellato, Francesca. (2018). “The Business History of the Preindustrial World: Towards a Comparative Historical Analysis.” Business History 61 (2): 225–59. doi: https://doi.org/10.1080/00076791.2018.1426750.

Het Woordenboek der Nederlandsche Tal (WNT) dictionary for early modern Dutch by the Instituut voor de Nederlandse Taal (Institute for the Dutch Language), entries for the terms “pijpaarde”, “droggoed", “crap/krap”, “panne” and “vetwaar”.

Janardhana Rao, Peluru, Nageswara Rao, Kunjam, Gokuruboyina, Sitaratnam, Neeraja, Kondamudi Naga. (2024). An Efficient Methodology for Identifying the Similarity Between Languages with Levenshtein Distance. In: Kumar, A., Mozar, S. (eds) Proceedings of the 6th International Conference on Communications and Cyber Physical Engineering. ICCCE 2024. Lecture Notes in Electrical Engineering, vol 1096. Springer, Singapore. https://doi.org/10.1007/978-981-99-7137-4_15

Jansson, Karin Hassan, Jonas Lindström, and Maria Ågren, 'Conclusion: On the Threshold of Modern Society', in Maria Ågren (ed.), Gender, Work, and the Transition to Modernity in Northwestern Europe, 1720–1880 (Oxford, 2025; online edn, Oxford Academic, 22 Jan. 2025), https://doi.org/10.1093/9780198934325.003.0010, accessed 11 Feb. 2026.

La Mela, Matti, Tamper, Minna and Kettunen, Kimmo (2019). Finding Nineteenth-century Berry Spots: Recognizing and Linking Place Names in a Historical Newspaper Berry-picking Corpus. In C. Navarretta, M. Agirrezabal & B. Maegaard (eds), Digital Humanities in the Nordic Countries: Proceedings of the Digital Humanities in the Nordic Countries 4th Conference. CEUR Workshop Proceedings, vol. 2364, CEUR-WS.org, Aachen, pp. 295-307

Moring, Beatrice (2025). The Strong Widows of the North — Women and Economic Activity in Nineteenth Century Helsinki. Journal of Family History, 50(4), 339-358. https://doi.org/10.1177/03631990241309840 (Original work published 2025)

Scheltjens, Werner (2023). “The feasibility of machine-learning based workflows for editing serial historical sources. First results of the Schenkenschans customs registers project”. In Ulrike Henny Krahmer (ed.), Machine learning and data mining for digital scholarly editions, Norderstedt.

Uppsala University, Gender and Work (GaW) research unit website: https://www.uu.se/en/research/gender-and-work. Accessed on 4 February 2026.

Wilson, John F., Jones, Ian G., Toms, Steven, Tilba, Anna, Buchnea, Emily, & Wong, Nicholas. (2022). Business History: A Research Overview (1st ed.). Routledge. https://doi.org/10.4324/9780429449536

Session 5B — 08:30–10:30

08:30–08:50 **Scattered throughout (Digital) Libraries: The Case of Old Writings in Baltic languages**
Ernesta Kazakėnaitė

08:50–09:10 **Lifting Local Treasures: Building Digital Access to Paper Sources from Danish Local Archives using Collaborative, Non-Profit Infrastructure**
Mia Gulvad Jørgensen, Andy Stauder

09:10–09:40 **Close reading, automation and cultural memory. Experiments with literary summarization through LLMs (Abstract)**
Alexander Conroy, Kirstine Nielsen Degn, Jens Bjerring-Hansen, Ali Al-Laith, Daniel Hershcovich, Matthew Wilkens

09:40–10:10 **Annotation Fever! An Interdisciplinary Experiment Exploring the Image to Come in Generative AI**
Klara Källström, Thobias Fäldt, Bernard Geoghegan, Mats Fridlund

10:10–10:30 **Where the Machine Looks Away**
Teodora Crisan-Matcaboja

08:30–08:50 *SHORT PAPER* **[93]**

Scattered throughout (Digital) Libraries: The Case of Old Writings in Baltic languages

Ernesta Kazakėnaitė
Vilnius University, Lithuania

Keywords: *16th-17th Old Writings, Baltic languages, digital collections, bibliography*

Due to historical circumstances, writings in the Baltic languages have dispersed across numerous libraries and archives, not only within the Baltic countries themselves but also far beyond their borders. The political, cultural, and religious transformations of the 16th and 17th centuries – together with migration, trade, and scholarly exchange – contributed to a remarkable diffusion of printed works. As a result, early books in Latvian, Lithuanian, and Old Prussian can now be found in unexpected locations across Europe and even overseas. For instance, Latvian books and pamphlets reached Denmark, Sweden, Estonia, and Germany, often through private collections, and later found their way into national and university libraries. A similar story can be told about Lithuanian and Prussian works, which were printed in various European printing centers and then distributed widely through church networks, academic institutions, and the activities of individual collectors.

In the 20th century, bibliographers and historians made significant efforts to record and catalogue these materials. Their work, published in bibliographies and catalogues (see Binkytė et al.; Daugaravičienė et al.; SLV), became the foundation for research in Baltic linguistics, literature, and cultural history. However, the transition to the digital age has brought new challenges and opportunities. Today, we face the task of reassembling this dispersed written heritage once again – this time in a digital environment. The question is no longer only *where* these books are physically located, but *which* of them have already been digitized, in what quality, and with what level of accessibility.

This task is far from simple. European libraries and heritage institutions are undergoing rapid and continuous digitization processes, but their pace, priorities, and metadata standards vary greatly. It is therefore difficult, and sometimes impossible, to know when a specific work will be digitized or how it will be made available to the public. In addition, digital collections are often scattered across different platforms, portals, and repositories, each with its own search interface and descriptive conventions. Researchers interested in Baltic studies are thus faced with a fragmented digital landscape that requires considerable time and expertise to navigate.

In order to address these difficulties and to facilitate access to early Baltic writings, a new initiative has been launched to compile a comprehensive bibliography of digital publications from the 16th and 17th centuries. This bibliography will bring together bibliographic data and links to digital copies of books in Baltic langugaes available in different libraries and archives, both within and beyond the Baltic region. The goal is to create a single, reliable, and regularly updated resource that enables researchers, linguists, and cultural historians to locate and consult these rare materials more easily.

In this presentation, I will discuss the ongoing work of identifying and cataloguing such books, as well as the development of the accompanying database (see https://www.kazakenaite.com/old-lithuanian). Special attention will be given to the methodology of data collection, the criteria for inclusion, and the variety of digital resources currently in use. I will also highlight the challenges and unexpected discoveries that have emerged during the project, including cases where books previously thought lost have resurfaced in digital form. Finally, I will consider how collaboration among libraries, scholars, and digital humanities initiatives can help sustain and expand this effort, ensuring that the written heritage of the Baltic languages continues to be accessible for future generations.

References:

Binkytė Eleonora... [et al.] (eds.). 1969. *Lietuvos TSR bibliografija, Serija A: Knygos lietuvių kalba, T.1 1547–1861.* Vilnius: Mintis.

Daugaravičienė Alma... [et al.] (eds.). 1990. *Knygos lietuvių kalba. T. 1. 1547-1861: papildymai*. Vilnius: Mintis.

SLV = Seniespiedumi latviešu valodā, 1525–1855: kopkatalogs = Die älteren Drucke in lettischer Sprache 1525–1855: Gesamtkatalog. By S. Šiško vadībā, zin. red. A. Apīnis. Rīga: Latvijas Nacionālā bibliotēka, 1999

References

Binkytė Eleonora... [et al.] (eds.). 1969. Lietuvos TSR bibliografija, Serija A: Knygos lietuvių kalba, T.1 1547–1861. Vilnius: Mintis.

Daugaravičienė Alma... [et al.] (eds.). 1990. Knygos lietuvių kalba. T. 1. 1547-1861: papildymai. Vilnius: Mintis.

SLV = Seniespiedumi latviešu valodā, 1525–1855: kopkatalogs = Die älteren Drucke in lettischer Sprache 1525–1855: Gesamtkatalog. By S. Šiško vadībā, zin. red. A. Apīnis. Rīga: Latvijas Nacionālā bibliotēka, 1999

08:50–09:10 *SHORT PAPER* **[94]**

Lifting Local Treasures: Building Digital Access to Paper Sources from Danish Local Archives using Collaborative, Non-Profit Infrastructure

Mia Gulvad Jørgensen[1], Andy Stauder[2]

[1] *Sammenslutningen af Lokalarkiver*

[2] *READ-COOP SCE*

Keywords: *Keywords: local archives; Denmark; digitisation; HTR/OCR; cooperative infrastructure; decentralisation; metadata interoperability; volunteer engagement; democracy.*

Keywords: local archives; Denmark; digitisation; HTR/OCR; cooperative infrastructure; decentralisation; metadata interoperability; volunteer engagement; democracy.

Background & Aim: This project aims to make previously inaccessible local historical sources in Denmark available for research and public use. Hundreds of local and town archives hold photographs, association minutes, and private papers vital for community history, democratic accountability, and practical needs (e.g., property boundaries, climate documentation). However, many collections remain invisible due to limited budgets, volunteer-run operations, and inconsistent digitisation. The project proposes a pragmatic approach by combining the national association **Sammenslutningen af Danske Lokalarkiver (SLA)**, the cooperative, non-profit-oriented digital infrastructure of **READ-COOP SCE** and its **Transkribus** platform, and workflows tailored for very small archives. Transkribus, developed from EU-funded research, offers text recognition (HTR/OCR), semantic extraction, layout analysis, and APIs for integration (READ-COOP SCE 2025a).

Research Questions:

1. How can a volunteer-heavy, resource-constrained network coordinate digitisation, transcription, and description at scale without imposing technical burdens?
2. What minimal, reusable building blocks (training, policies, tooling) move small archives from box-level backlogs to item-level public discovery?
3. What are the sustainability and governance advantages/risks of cooperative, non-profit infrastructure compared to commercial platforms, especially for decentralised cultural heritage?

Context & Materials: SLA's ecosystem feeds **arkiv.dk**, a national portal built on the **Arkibas** cataloguing system (Arkibas ApS 2025), aggregating 580+ local archive collections. Millions of records are queryable, but much material remains uncatalogued or undigitised. Target materials include property/fire marshal records, council minutes, school boards, unions, associations, clubs, alms-houses, and hospitals.

Motivation (democracy, decentralisation, equality): Direct, decentralised access to primary sources strengthens local democracy, reduces centre-periphery inequalities, and enables evidence use in current administrative or environmental cases. Cooperative infrastructure aligns with community governance and public-interest stewardship, as documented in recent READ-COOP/Transkribus analyses (Terras et al. 2025).

Methods & Workflow (for "even the smallest archive"): The pipeline is designed for feasibility (2–3 volunteer hours/week) and low cost:

Selection & intake: A triage rubric prioritises "high-impact/low-effort" series and flags rights-sensitive material à **Digitisation:** Deploy **ScanTent** smartphone kits (READ-COOP SCE 2025b) for affordable, quality scanning. Volunteers receive checklists, safety guidance, and training. à **Text, layout, and named-entity recognition:** A core team of up to 35 volunteers, guided by SLA staff and READ-COOP, uses Transkribus models, fine-tuning for Danish/collection-specific needs. Automatic segmentation and annotation extract key fields for description à **Access to infrastructure:** SLA's large-scale membership in READ-COOP would enable co-ownership, unlimited processing for a set number of accounts, co-determination rights, privileged information, and reduced financial contributions. This is feasible with a modest one-time and annual fee per archive (two-to-three-figure € amounts). à **Collaborative**

correction: Volunteers validate key entities via browser-based correction; periodic checks yield CER/WER metrics to calibrate quality. à **Description & interoperability:** A minimal crosswalk maps local fields to a harmonised schema (identifier, collection, date, place, rights, persons), with export profiles for Arkibas/arkiv.dk to avoid duplicate entry (based on Andresen 2007). à **Search & publication:** Items and transcriptions flow to a public site with provenance, rights statements, and uncertainty cues; where feasible, surfaced back into arkiv.dk.

A parallel part of the pipeline is **integration:** Exploring API-level integration between Transkribus and Arkibas to reduce “ye-another-tool” friction.

Current Stage: Work is underway on an inventory of document types and bottlenecks, and concrete use-cases (democratic fact-checking, property disputes, environmental history), alongside legal and ethical screening. Funding models are being prepared, including public grants and volunteer time. Local archives are being integrated into the READ-COOP community, and initial dry runs confirm a no-code pipeline from capture through extraction to correction and export. This aligns with recent peer-reviewed work on READ-COOP’s community governance (Terras et. al 2025).

Early Observations & Novelty: Capacity is the main bottleneck - equipment and volunteer time matter more than algorithmic performance. Templated micro-workflows are crucial. Cooperative, no-code tooling reduces lock-in and suits non-technical contributors (Terras et al., 2025). No other national initiatives were found that combine AI-enabled processing with volunteer-led workflows across hundreds of small archives on minimal budgets. READ-COOP self-reported eco-responsibility includes 100% renewable energy.

Planned Next Steps (2026):

Funding access: Evaluate legal/financial frameworks and prepare contracts for SLA and READ-COOP community interface.

Core volunteer team: Survey availability and willingness for the core team.

Volunteer training: Three-evening curriculum: scanning, training-data creation/correction, metadata & GDPR compliance.

Scan flow: Survey scanning capabilities, provide decision trees for equipment, documentation, and self-help resources for uploading, and assign trained volunteers to advise archives.

Pilot: Execute an end-to-end “small stack” on ~5k pages; report accuracy and evaluate need for dedicated model training.

Reusable assets: Open capture checklists, metadata crosswalk, GDPR decision tree, Arkibas/arkiv.dk export profiles.

Public search beta: Launch a minimal site with collection pages, rights, and feedback; plan syndication to arkiv.dk.

Integration options: Feasibility note on Transkribus ↔ Arkibas via metagrapho API; UI sketches for one-click send to Arkibas.

Partnerships: Convene a conference with other Danish archival institutions for collaboration.

Feedback Sought:

Reaching all archives: What outreach, micro-grants, and training formats entice both low-IT and professionalised archives?

Quality vs. effort: What is a defensible target for page-level accuracy under volunteer QA, and how should uncertainty be communicated?

Integration path: Is API-level integration the right priority, or should export/import profiles and training come first?

Ethics & GDPR: What selection/redaction rules for living persons are adequate and auditable? What complaint/withdrawal pathways are expected?

Decentralised equity: How to measure reduction in centre–periphery inequalities (KPIs beyond page counts)?

Comparators: Examples of low-budget, national-network digitisation with AI extraction for benchmarking.

Openness & ethics: Digitised items and transcriptions will be publicly searchable, with clear provenance and rights statements. Sensitive material will be screened and redacted/excluded as needed. Infrastructure choices favour cooperative governance and community ownership. Sustainability claims will be presented as self-reported by READ-COOP.

Expected Outputs by DHNB 2026:

Pilot setup

Public beta of the search site with ~5k pages

Open templates: capture checklists, metadata crosswalk, GDPR decision tree

References

Andresen, Leif (2007) Common presentation of data from archives, libraries and museums in Denmark. Danish Library Agency, October 2007. <https://slks.dk/fileadmin/user_upload/0_SLKS/Dokumenter/Biblioteker/Standarder/ABM/Common_presentation_of_data_from_archives_libraries_and_museums_Denmark.pdf>

Arkibas ApS (2025). “About” section of the Arkibas.dk portal. <https://arkiv.dk/en/om>

READ-COOP SCE (2025a). Website about the customisable, community-owned AI platform “Transkribus”. <https://www.transkribus.org/>

READ-COOP SCE (2025b). Website about the ScanTent low-cost scanning aid. <https://www.transkribus.org/scantent >

Terras, Melissa et al. (2025). Open Research Europe (Article 5-16, 2025). <https://open-research-europe.ec.europa.eu/articles/5-16>

09:10–09:40 *LONG PAPER* **[95]**

Close reading, automation and cultural memory. Experiments with literary summarization through LLMs (Abstract)

Alexander Conroy[1], Kirstine Nielsen Degn[1], Jens Bjerring-Hansen[1], Ali Al-Laith[1], Daniel Hershcovich[1], Matthew Wilkens[2]

[1] *The University of Copenhagen, Denmark*
[2] *Cornell University, USA*

Keywords: *Computational Literary Studies, Literary canons and archives, Text summarization, Literary Historiography*

1. Introduction

This paper addresses the basic literary historiographical practice of summarization and, particularly, how Large Language Models (LLMs) and generative AI can assist in performing it. The ability to automatically summarize literary works holds immense potential for scholars seeking to orient themselves within the vast number of texts lost in abundance, often referred to as “the great unread” (Cohen 1999, 23). In this context, where paratextual and bibliographic information about many works is scarce, AI-based summarization can provide valuable overviews that facilitate literary explorations and, thereby, social expansions of our empirical object. From a historical perspective, summarization can be viewed as a foundational humanistic practice, with roots dating back to early modern traditions of epitomization and cataloging, which ease access to literary and scientific information. Beyond these scholarly utilities, book summarization through LLMs also speaks to a broader cultural and social function, as it echoes a central question that connects readers and reading communities alike: what is the book about?

2. Theory

Literary summaries are often perceived as instruments of scientific control (Kondrup 2019, 23–24) or as pedagogical and analytical steppingstones (Vendler 2002, 126). As such, summarization, the brief going-over of key narrative elements, is both a checkpoint for foundational textual understanding and a point of departure for diverse theoretical and ideological interpretations. In this light, summarization operates at the intersection of factuality and selectivity, as it must convey verifiable information while foregrounding the most salient features of a work. We propose a framework that centers on five key features – fictional time, spatial setting, main characters, key plot events, and central themes – as the basis for assessing and generating basic literary summaries. Such summaries must balance precision with critical attention, reducing the narrative complexity into a form that remains both informative and indicative.

We approach this task computationally through LLMs in a twofold manner. First, we ask how – and to what extent – it is possible to generate summaries that capture a literary work's essential narrative structure while retaining only a minimal yet necessary interpretive layer. Second, we explore how summaries might be generated as interpretive companions to the specific analytical needs of literary scholars. The latter approach resonates with the idea of guiding or tuning summaries toward distinct interpretive and theoretical agendas (Kramnick 2023, 69–70). Whereas the first approach treats summarization as both an instrument for demonstrating basic textual comprehension and an analytical starting point, the second is more deeply entangled in interpretation itself.

Importantly, these automated approaches demand scholarly scrutiny: when relying on the products of an increasingly closed off tech industry, there's a risk of compromising core scientific values such as transparency and reproducibility. Such considerations are not merely peripheral but constitute central concerns of this paper as it examines both the possibilities and the methodological challenges of employing LLMs for literary summarization.

3. Methods

The automatic summarization of novels and other long-form narratives represents a significant frontier for computational text analysis. Literary narratives differ sharply from the kinds of texts machines summarize well – like news articles – because their plots are complex, their characters develop over time, and their narration shifts across modes and perspectives. Advances in summarization techniques could open new avenues for exploring large collections of literary texts – especially those that remain understudied, marginalized, or otherwise situated outside the traditional literary canon.

In this paper we will examine the capacity of LLMs to summarize Danish language novels from the late nineteenth century. Adopting a zero-shot approach – that is, prompting models to summarize novels without any fine-tuning or presenting examples of summaries – makes it possible to evaluate how well current systems can identify and reproduce core narrative structures using only their general language understanding. In contrast to few-shot or fine-tuned setups, which rely on curated examples, the zero-shot configuration offers a clearer view of the models' baseline ability to process and condense complex literary texts. A select group of models are prompted to produce Danish summaries of a set of novels, with instructions emphasizing factual narrative elements, while keeping interpretive commentary and stylistic elaboration at a minimum.

To evaluate the literary adequacy of the generated summaries, we have developed a new benchmarking framework grounded in baseline features of fictional, narrative texts: time, place, characters, plot, and themes. Human annotators will assess each summary according to its coverage and accuracy in these five dimensions as well as criteria more related to the linguistic quality of the model outputs (e.g., fluency and coherence). This scheme is developed to provide a general framework for summarizing novels, but it can readily be tailored to address specialized analytical needs or particular literary features.

Following this evaluation setup, we will conduct a series of prompt engineering experiments testing how instruction design, role framing, and linguistic phrasing influence the precision and coherence of model outputs. The purpose of these experiments is not to systematically test all possible prompt configurations – a task that would be both complex and methodologically elusive – but rather to identify which strategies appear to produce the most coherent and factually grounded summaries within this literary context.

Finally, to probe the limits of model knowledge and cultural inference, we will conduct a metadata-only experiment. Here, models are asked to summarize novels without access to the text, receiving only the author name and title. The test contrasts one highly canonical work (e.g., Niels Lyhne by J. P. Jacobsen)

with one obscure text (e.g., Bjergmandens Ring by the unidentified pseudonym Claudius Albertius), illustrating how models rely on cultural priors and canonical knowledge when textual evidence is unavailable.

4. Discussion

This paper brings together two principled and interrelated discussions, both grounded in qualitative analyses of model outputs.

First, it offers a theoretical discussion of what constitutes a good summary. By examining automatically generated summaries and comparing them with earlier summaries produced in scholarly contexts, the paper explores the metatheoretical framework of the 'basic summary' as opposed to summaries shaped by specific interpretive or theoretical interests.

Second, it presents a conceptual discussion of the challenges involved in establishing a framework capable of summarizing both well-known canonical novels and completely forgotten works of popular fiction. This discussion is also anchored in specific cases – for instance, Hunger (Sult) by Knut Hamsun, a proto-modernist and early surrealist masterpiece that helped shift the novel toward modern aesthetics, psychological depth, and ontological fragmentation (Sørensen 2015), contrasted with one of the many novels from the same period categorized under the now entirely obscure genre original romance (Original romantisk Fortælling).

The range of challenges connected to these discussions extends from the tangible such as the models' ability to handle context length – we know canonical novels from the period are typically short, while popular fiction tends to be much longer (Bjerring-Hansen and Rasmussen 2023) – to the more abstract: can models process literary complexity and conventionality equally well? By addressing such issues, the paper not only proposes an analytical strategy for literary summarization but also contributes to a broader understanding of how LLMs engage with the formal and cultural dimensions of literary history.

References

Bjerring-Hansen, Jens, and Sebastian Ørtoft Rasmussen. 2023. "Litteratursociologi og kvantitative litteraturstudier: Den historiske roman i det moderne gennembrud som case." Passage – Tidsskrift for litteratur og kritik 38 (89): 171–189.

Cohen, Margaret. 1999. The Sentimental Education of the Novel. Princeton: Princeton University Press.

Kondrup, Johnny. 2019. Bjergtaget. Illusion og forførelse fra Søren Kierkegaard til Karen Blixen. Skive: Forlaget Wunderbuch.

Kramnick, Jonathan. 2023. Criticism and Truth: On Method in Literary Studies. Chicago: University of Chicago Press.

Sørensen, Peer E. 2015. Hamsuns sultekunstner. Aarhus: Aarhus University Press.

Vendler, Helen. 2002. Poems, Poets, Poetry: An Introduction and Anthology. 2nd ed. Boston: Bedford Books of St. Martin's Press.

09:40–10:10 *LONG PAPER* **[96]**

Annotation Fever! An Interdisciplinary Experiment Exploring the Image to Come in Generative AI

Klara Källström[1,2], Thobias Fäldt[2], Bernard Geoghegan[1], Mats Fridlund[1]

[1] *University of Gothenburg, Sweden*

[2] *FG2, Sweden*

Keywords: *machine learning, annotation, artistic research, vernacular photography, iconology*

The following presentation describes an ongoing interdisciplinary project situated across digital humanities and artistic research, and across GLAM and academia. It focuses on developing an approach that critically explores the current and future state of AI by studying an application practice central to generative AI in the form of image annotation. The presentation gives an overview of the background to the project, the hands-on and exhibitionary work carried out so far, its conceptual foundations in critical theory and digital studies, as well as planned digital development work to take place in future stages, when the project aim to develop a prototype of a generative AI image model based on the more diverse and varied human categorizations of photographic images than the

commercial AI image models currently available. At its core, the project asks how human and artificial acts of seeing and naming might be understood and reimagined through the use of experiments that explore collective and situated practices that unfold across artistic and technical domains. Through these situated experiments it positions image annotation as both a practical and philosophical inquiry into the infrastructures that sustain machine vision, foregrounding the embodied and interpretive dimensions of human perception. The project thereby links technical procedures of labeling to broader aesthetic and epistemological questions about how meaning is distributed, shared, transformed, and sustained in computational and digital environments.

Annotation Fever! unfolds through a series of exhibitions and workshops developed collaboratively by artists Klara Källström and Thobias Fäldt, media theorist Bernard Dionysius Geoghegan, and GRIDH (Gothenburg Research Infrastructure for Digital Humanities, University of Gothenburg). This collaboration was initiated during Spring 2025 through a chance encounter of Källström and Fridlund where Källström asked if it would be possible to develop an AI image generation model trained on previously largely unexplored image corpora in the form of two digital photographic archives. This initial contact was followed by a period of discussions between Källström, Fäldt, Geoghgan and GRIDH research engineers using their collective expertise to explore the affordances of generative AI, digital photographs, computer vision and digital annotation tools to investigate the possibility of a joint project that would explore the built-in conceptual and political choices and biases of existing and possible AI image generation models. This led to the first instantiation of the larger project in the form of a pilot project to generate training data for a possible future AI image generation model through a succession of manual image annotation workshops staged in September 2025 during the Gothenburg International Biennale for Contemporary Art (GIBCA) and continuing in December at Camera Austria in Graz. At these workshops some hundred non-expert users (exhibition visitors and students) were manually annotating a selection of images from the two photographic archives and thus the project turned the manual practice of image annotation from a repetitive and mundane task usually taking place in commercial image annotation into an aesthetic and epistemic experiment, an evolving collective practice of attention and critique.

The basis for the project is the exploration of two extensive digital photographic archives that together provide the foundation for developing a AI image generation model in future stages: Thomas Sauvin's *Beijing Silvermine*, a vast collection of about a million vernacular Chinese photographic negatives rescued from destruction, and Källström and Fäldt's evolving photographic corpus, which engages questions of representation and the politics of the image. Together, these archives will be used to probe collaboration at the dataset level, engaging with the latent photographic and conceptual space of neural networks and AI image generation models. *Annotation Fever!* stages an encounter between art, digital humanities, machine vision and human perception, exploring through experiments with non-professional users how the taken for granted and black-boxed data of generative AI might be made perceptible through new ways of probing and imaging its imaging production.

Within the project, we inquire into how digital-humanities methodologies might adapt to attend to this moment of becoming. More specifically, we explore how dimensions of annotation can be understood through conditions in which recognition emerges, and whether the classificatory categories upon which generative AI depends could be otherwise, examining how alternative epistemic grammars or world models might be imagined or enacted through digital humanities practice.

In this sense, *Annotation Fever!* situates itself within the DHNB 2026 theme *Lost in Abundance: Encounters with the Non-Canonical*. Generative AI's visuality arises through mainly hidden and unexplored vast commercial infrastructures of human labour and perception, as can be seen in the number of Human Intelligence Task (HIT) jobs listed on Amazon Mechanical Turk. The annotator, often overlooked in narratives of creativity, appears here as a figure through which the non-canonical takes shape in digital culture. Through its use of annotation workshops where non-digital experts manually annotate images, the project transforms annotation from an isolated task into a participatory and reflective process, and thus *Annotation Fever!* reclaims annotation as a site of negotiation and co-creation. Each manual mark, consent, or omission becomes a practical micropolitical gesture, determining what may circulate and what must remain outside the computational archive of AI image generation models.

In today's generative AI, the photographic is everywhere and nowhere at once. Photography's histories, conventions, and archives form the basis of training large datasets, which in turn shape the statistical systems through which new images are read and generated. As a dataset becomes computationally operative, acts of recognition translate visual experience into the modularity of machine-readable

language. Annotation lies at the heart of this process. Central to this inquiry is the question of which visible and hidden categories and conceptualizations mutually structure human and machine learning, including large language models (LLMs). The principles through which AI datasets are organized seek to emulate linguistic order, producing what Steyerl (2023) calls “mean images”: statistically averaged, rule-based composites that reinforce normative visibility. These operations contribute to what Amoore et al. (2024) describe as “a world model,” a computational epistemology that arranges the world through probabilistic inference and grammar-like rules. *Annotation Fever!* renders these structures perceptible by tracing the modularity of language that underpins data annotation, turning attention to what unfolds within this logic.

This condition is also historical. As Benjamin (1936/2008) observed, human perception is shaped through history, a collective medium in which nature, seeing, and knowing are organized. The photographic archives mobilized by *Annotation Fever!* enact this idea, suggesting that each image carries the sediment of collective perception. By reactivating *Beijing Silvermine* and Källström and Fäldt’s corpus through acts of manual annotation in workshops with non-expert users, the project examines how the past continues to be inscribed within evolving regimes of artificial and human visibility.

Considering who holds the right to archive and in what ways, and following the logics that sustain the compulsion to turn everything into archival items, Derrida’s (1995) *Archive Fever* attends to what emerges. *Annotation Fever!* extends this gesture through the collective process of annotation as co-creation. By keeping their archives open to revision, Källström, Fäldt, Sauvin, and Geoghegan approach the archive as a field of micropolitics in which acts of labeling and inscription are continuously negotiated. The exhibition and workshops make these negotiations perceptible, attending to what may be withheld, erased, or left unmarked. The resulting images invite reflection of authorship and ownership—how the annotated image relates to its makers, and whether it is to be circulated, donated for further training, or withheld from computational afterlives. To annotate a model is not only to segment an image into machine-legible units but to shape the terms of vision itself, marking what tomorrow’s collective perception will become. This project is an attempt to explore how we and our digital machines could have seen and could see otherwise. In the past, in the present and in the future.

References

Amoore, L., Campolo, A., Jacobsen, B., & Rella, L. (2024). *A world model: On the political logics of generative AI*. *Political Geography, 113*, Article 103134. https://doi.org/10.1016/j.polgeo.2024.103134

Benjamin, W. (2008). *The work of art in the age of its technological reproducibility, and other writings on media* (M. W. Jennings, B. Doherty, & T. Y. Levin, Eds.). Harvard University Press. (Original work published 1936)

Derrida, J. (1995). *Archive fever: A Freudian* impression (E. Prenowitz, Trans.). University of Chicago Press.

Steyerl, H. (2023). Mean images. *New Left Review*, 140/141, 101–118.

10:10–10:30 *SHORT PAPER* **[97]**

Where the Machine Looks Away

Teodora Crisan-Matcaboja

Babes-Bolyai University, Romania

Keywords: *Trauma theory, Algorithmic memory, Generative AI, Ethics of representation, Digital infrastructures*

Image-generation systems promise endless visual abundance, yet their limits surface the moment they are asked to picture pain. When prompted with wars, displacement, or historical atrocity, models such as Stable Diffusion XL or Midjourney frequently avoid direct depiction through refusal, sanitization, or representational displacement, producing images that reformat harm into safer, legible visual forms. This paper investigates how such algorithmic evasions form part of a wider cultural pattern: the translation of trauma into data—a problem long theorized in critical debates on the ethical and aesthetic

risks of rendering suffering legible through formal systems (Adorno 1981)—and the disappearance of what cannot be formatted.

The study asks how generative AI systems handle traumatic or ethically charged prompts, what forms of visual or textual refusal they produce, and how those refusals might be interpreted through trauma theory. Building on trauma studies (Caruth 1996; Kaplan 2008; LaCapra 2001) and critical analyses of AI infrastructures (Crawford 2021; Chun 2021; Zylinska 2020), I argue that generative systems reproduce, in computational form, strategies long associated with the psyche's handling of trauma (Hayles 2017; Hui 2016). They enact collapse, reducing multiple and contradictory affects into coherent, viewable form, while traces of what cannot be represented return as distortion, refusal, or silence. Together with superposition (the coexistence of incompatible states) and leak (the remainder that escapes containment), these terms form a structural grammar for reading how trauma circulates through algorithmic media. They are used not as metaphors or claims about machine cognition, but as analytic operators adapted from trauma theory to describe how generative systems format, contain, or fail to integrate ethically charged content. Rather than treating safety filters as external constraints, the paper reads moderation, refusal, and aesthetic softening as integral to the representational logic of generative systems, where ethical governance and technical mediation actively shape what becomes visible or sayable (Crawford 2021; Chun 2021; Zylinska 2020; Benjamin 1936).

The material consists of fifty text-to-image prompts derived from public war-archive descriptors. Images were generated using Stable Diffusion XL via the Clipdrop web interface, with default safety filters enabled, and each run was documented with respect to both output and refusal. The prompt set was thematically constructed rather than randomly sampled, allowing for controlled comparison across historically and ethically charged scenes. Outputs and moderation responses were coded using predefined criteria as render (direct depiction), sanitize (symbolic substitution or aesthetic softening), or refuse (explicit moderation), with ambiguous cases documented and treated systematically. While the study centers on a single model, it treats SDXL as a representative instance of broader generative image systems shaped by shared data regimes and moderation infrastructures, as documented in critical accounts of AI as cultural infrastructure (Crawford 2021; Chun 2021).

The methodology combines descriptive statistics with qualitative visual analysis to identify recurrent representational patterns. Repeated behaviors—symbolic substitution, aesthetic softening, or blank refusal screens—are read not as technical faults but as cultural symptoms. Preliminary findings indicate that over half of the prompts were resolved through symbolic substitution, ontological displacement, or epistemic reframing rather than direct depiction, indicating that generative systems manage traumatic content by converting harm into administratively and culturally legible forms rather than simply suppressing it. In this framing, the non-canonical image is not a glitch in the system but the traumatic remainder that resists integration into data infrastructures.

Methodologically, the paper brings psychoanalytic and digital-humanities vocabularies into dialogue. Where digital-heritage research often speaks of bias or missing data, trauma theory offers a deeper logic of absence, repetition, and deferred return. Mapping superposition–collapse–leak onto dataset curation, image generation, and moderation reveals how AI systems aestheticize, suppress, and occasionally betray traumatic memory. In their gestures of protection and refusal, generative systems expose the boundary where computation meets cultural memory. Read through trauma theory, algorithmic "safety" emerges not as neutral precaution but as a record of repression itself—a sign that even our machines inherit the desire to forget.

Beyond documenting algorithmic behavior, the study contributes to ongoing debates in digital humanities concerning ethics, affect, and interpretive labor in computational environments. By shifting attention from what models generate to what they exclude, it reframes generative datasets and digital archives as active sites of cultural negotiation. The analysis highlights how moderation filters, aesthetic defaults, and interface design collectively enact a politics of legibility, deciding which forms of pain are permitted to appear and which remain unrendered. Ultimately, the paper argues that reading machine refusals through trauma theory opens a path toward a more reflexive and ethically attentive digital humanities—one that treats absence and silence not as voids, but as meaningful evidence of how culture and computation remember differently.

References

Adorno, Theodor W. Prisms. Translated by Samuel and Shierry Weber. Cambridge, MA: MIT Press, 1981.

Benjamin, Walter. The Work of Art in the Age of Its Technological Reproducibility. Translated by Edmund Jephcott and Harry Zohn. Cambridge, MA: Harvard University Press, 2008. (Original essay published 1936.)

Caruth, Cathy. Unclaimed Experience: Trauma, Narrative, and History. Baltimore: Johns Hopkins University Press, 1995.

Chun, Wendy Hui Kyong. Discriminating Data: Correlation, Neighborhoods, and the New Politics of Recognition. Cambridge, MA: MIT Press, 2021.

Crawford, Kate. Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence. New Haven: Yale University Press, 2021.

Hayles, N. Katherine. Unthought: The Power of the Cognitive Nonconscious. Chicago: University of Chicago Press, 2017.

Hui, Yuk. On the Existence of Digital Objects. Minneapolis: University of Minnesota Press, 2016.

Kaplan, E. Ann. Trauma Culture: The Politics of Terror and Loss in Media and Literature. New Brunswick, NJ: Rutgers University Press, 2005.

LaCapra, Dominick. Writing History, Writing Trauma. Baltimore: Johns Hopkins University Press, 2014.

Zylinska, Joanna. AI Art: Machine Visions and Warped Dreams. London: Open Humanities Press, 2020.