DHNB 2025

PROGRAMME

BOOK OF ABSTRACTS

Digital Dreams and Practices

Digital Humanities in Nordic and Baltic Countries
9th Conference

VENUE: ESTONIAN NATIONAL MUSEUM

TARTU, ESTONIA

5-7/03/2025

DHNB

visit estonia

Co-funded by the European Union

Investing in your future

ORGANISERS:
Estonian Literary Museum (Mari Väina, Kadri Vider, Liisi Laineste, Olha Petrovych)
Estonian National Museum (Pille Runnel, Agnes Aljas, Piibe Nõmm)
Institute of the Estonian Language (Martin Eessalu)
Tallinn University (Andres Karjus)
University of Tartu (Liina Lindström, Joshua Wilbur)
in collaboration with Estonian Society for Digital Humanities

SUPPORTED BY:

TK215 Estonian Roots: Centre of Excellence for Transdisciplinary Studies on Ethnogenesis and Cultural Diversity (funded by the Ministry of Education and Research).

PRG1288 A Corpus-Based Approach to Folkloric Variation: Regional Styles, Thematic Networks, and Communicative Modes in Runosong Tradition (funded by the Estonian Research Council).

MOB3JD1218 Traditional Estonian and Ukrainian Folksongs: Comparative Corpus-Based Computational Analysis (funded by the Estonian Research Council).

Estonian Cultural Endowment (S07-24/0926)

KUM-TA43 Digital Cultural Heritage as a Resource for Social Development (funded by the Estonian Ministry of Culture)

The conference is co-funded by the European Union and Enterprise Estonia with funding of 27 064,31 euros (project No 2021-2027.1.03.24-0272).

visit estonia

Co-funded by
the European Union

Investing
in your future

CULTURAL ENDOWMENT OF ESTONIA

Editing and layout: Mari Väina, Olha Petrovych, Liisi Laineste

ISBN 978-9916-742-47-1 (pdf)

PROGRAMME

Date: Monday, 03/Mar/2025	
9:00am - 1:00pm **Oskar Kallas Auditorium**	**WS01: Decoding the Past, Digitizing the Future: Transkribus and (Digitized) Cultural Heritage I. Introductory Workshop** Location: **Oskar Kallas Auditorium** Session Chair: **C. Annemieke Romein**, READ-COOP SCE, Austria
9:00am - 1:00pm **Helmi Kurrik Auditorium**	**WS02: Innovations and New Interactions through Digital Cultural Heritage in GLAM Sector** Location: **Helmi Kurrik Auditorium** Session Chair: **Pille Runnel**, Estonian National Museum, Estonia
9:00am - 1:00pm **Aliise Moora Auditorium**	**WS03A: Explorations of the dynamics of cultural phenomena in text corpora I (30 min presentations & discussion)** Location: **Aliise Moora Auditorium** Session Chair: **Mark Mets**, Tallinn University, Estonia Session Chair: **Kadri Vider**, Estonian Literary Museum, Estonia Session Chair: **Taive Särg**, Estonian Literary Museum, Estonia Session Chair: **Katrine F. Baunvig**, Aarhus University, Denmark Session Chair: **Mari Väina**, Estonian Literary Museum, Estonia This session consists of six 30-min presentations and a slot for discussion.
2:00pm - 6:00pm **Aliise Moora Auditorium**	**WS03B: Explorations of the dynamics of cultural phenomena in text corpora II (60 min hands-on tutorials)** Location: **Aliise Moora Auditorium** Session Chair: **Mark Mets**, Tallinn University, Estonia Session Chair: **Kadri Vider**, Estonian Literary Museum, Estonia Session Chair: **Taive Särg**, Estonian Literary Museum, Estonia Session Chair: **Katrine F. Baunvig**, Aarhus University, Denmark Session Chair: **Mari Väina**, Estonian Literary Museum, Estonia This session consists of 3 hands-on tutorials of 60 minutes.
2:00pm - 6:00pm **Oskar Kallas Auditorium**	**WS04: Decoding the Past, Digitizing the Future: Transkribus and (Digitized) Cultural Heritage II. Advanced Workshop** Location: **Oskar Kallas Auditorium** Session Chair: **C. Annemieke Romein**, READ-COOP SCE, Austria
2:00pm - 6:00pm **Helmi Kurrik Auditorium**	**WS05: DH in Libraries, Archives and Museums DHNB working group meeting** Location: **Helmi Kurrik Auditorium** Session Chair: **Camilla Holm**, OsloMet - Oslo Metropolitan University, Norway Session Chair: **Olga Holownia**, IIPC, United States of America
Date: Tuesday, 04/Mar/2025	
9:00am - 1:00pm **Oskar Kallas Auditorium**	**WS06: From Data Cleanup to Linked Open Data: Hands-on with OpenRefine and Wikidata** Location: **Oskar Kallas Auditorium** Session Chair: **Alicia Fagerving**, Wikimedia Sverige, Sweden Session Chair: **Ida Nordlander**, Swedish Centre for Architecture and Design, Sweden
9:00am - 1:00pm	

Ilmari Manninen Auditorium	**WS07: Workshop on Digital Humanities and Social Sciences/Cultural Heritage (DHSS/DHCH) in Higher Education** Location: Ilmari Manninen Auditorium Session Chair: **Koraljka Golub**, Linnaeus University, Croatia Session Chair: **Marianne Ping Huang**, Aarhus Universitet, Denmark Session Chair: **Isto Huvila**, Uppsala University, Sweden Session Chair: **Jonas Ingvarsson**, Göteborgs universitet, Sweden Session Chair: **Ahmad Kamal**, Linnaeus University, Sweden Session Chair: **Olle Sköld**, Uppsala University, Sweden
9:00am - 1:00pm Aliise Moora Auditorium	**WS08: Tradition Archives Meet Digital Humanities II** Location: Aliise Moora Auditorium Session Chair: **Sanita Reinsone**, University of Latvia, Latvia Session Chair: **Kati Kallio**, Finnish Literature Society, Finland (Hybrid format)
9:00am - 1:00pm RaKuKe	**WS09: Web Archive Collections as Data** Location: RaKuKe Session Chair: **Olga Holownia**, IIPC, United States of America Session Chair: **Gustavo Candela**, University of alicante, Spain Session Chair: **Helena Byrne**, British Library, United Kingdom Session Chair: **Jon Carlstedt Tønnessen**, National Library of Norway, Norway Session Chair: **Anders Klindt Myrvoll**, Det Kgl. Bibliotek, Denmark
9:00am - 6:00pm Helmi Kurrik Auditorium	**WS10: Create and Showcase Your Digital Critical Edition: A Step-by-Step Guide with Digital Philology for Dummies (DPhD) and Edition Visualization Technology (EVT)** Location: Helmi Kurrik Auditorium Session Chair: **Chantal Pivetta**, Lund University (Sweden), Sweden Session Chair: **Renato Caenaro**, SilentWave SRLS, Italy Session Chair: **Roberto Rosselli Del Turco**, Università di Torino, Italy
2:00pm - 6:00pm RaKuKe	**S01: Doctoral Consortium** Location: RaKuKe
2:00pm - 6:00pm Oskar Kallas Auditorium	**WS11: How to create a Linked Open Data service and semantic portal for your own Cultural Heritage data** Location: Oskar Kallas Auditorium Session Chair: **Eero Hyvönen**, Aalto University and University of Helsinki (HELDIG), Finland Session Chair: **Jouni Tuominen**, University of Helsinki, Finland Session Chair: **Heikki Rantala**, Aalto University, Finland Session Chair: **Petri Leskinen**, Aalto University, Finland Session Chair: **Rafael Leal**, Aalto University, Finland Session Chair: **Annastiina Ahola**, Aalto University, Finland
2:00pm - 6:00pm Aliise Moora Auditorium	**WS12: Exploring, transforming and visualizing digital editions and corpora with visual and AI augmented workflows** Location: Aliise Moora Auditorium Session Chair: **Sasha Rudan**, LitTerra Foundation, Serbia Session Chair: **Eugenia Kelbert Rudan**, Instite of World Literature, Slovak Republic
2:00pm - 6:00pm Ilmari Manninen Auditorium	**WS13: How to Structure and Organize a National Digital Humanities Research Infrastructure: Realizing the Digital Dreams of Tomorrow** Location: Ilmari Manninen Auditorium Session Chair: **Edward Joseph Gray**, DARIAH-EU, France Session Chair: **Sanita Reinsone**, University of Latvia, Latvia Session Chair: **Koraljka Golub**, Linnaeus University, Croatia Session Chair: **Eiríkur Smári Sigurðarson**, University of Iceland, Iceland Session Chair: **Olga Holownia**, IIPC, United States of America Session Chair: **Mari Väina**, Estonian Literary Museum, Estonia
2:00pm - 6:00pm Maailmafilm	**WS14: Workshop on cultural heritage data mining** Location: Maailmafilm Session Chair: **Lars Kjær**, The Royal Danish Library, Denmark Session Chair: **Anders Klindt Myrvoll**, Det Kgl. Bibliotek, Denmark
Date: Wednesday, 05/Mar/2025	
9:00am - 11:00am	**Registration. Welcome coffee**

11:00am - 12:30pm Jakob Hurt’s Hall	**Plenary Session 1: Opening of the conference. Opening Keynote: Maciej Eder: Text Analysis Is Easy, Unless It Is Not: Reliability Issues in Measuring Textual Similarities** Location: Jakob Hurt’s Hall Session Chair: **Joshua Wilbur**, University of Tartu, Estonia
12:30pm - 1:30pm	**Lunch**
1:30pm - 3:30pm Ilmari Manninen Auditorium	**Panel 1: Show me your data! Visualization and interpretation of data from Cultural Heritage** Location: Ilmari Manninen Auditorium Session Chair: **Chantal Pivetta**, Lund University (Sweden), Sweden
1:30pm - 3:30pm Jakob Hurt’s Hall	**Session LP 01: Digital History** Location: Jakob Hurt’s Hall Session Chair: **Mats Fridlund**, University of Gothenburg, Sweden, Sweden
1:30pm - 3:30pm Helmi Kurrik Auditorium	**Session LP 02: Past and Present in the Digital Age** Location: Helmi Kurrik Auditorium Session Chair: **Sanita Reinsone**, University of Latvia, Latvia
1:30pm - 3:30pm Aliise Moora Auditorium	**Session LP 03: Digital Art, Media & Data Reimagined** Location: Aliise Moora Auditorium
3:30pm - 4:00pm	**Coffee Break**
4:00pm - 5:40pm Jakob Hurt’s Hall	**Session SP 01: AI, Archives and Digital Heritage** Location: Jakob Hurt’s Hall Session Chair: **Pille Runnel**, Estonian National Museum, Estonia
4:00pm - 5:40pm Helmi Kurrik Auditorium	**Session SP 02: AI and Text Analysis** Location: Helmi Kurrik Auditorium Session Chair: **Eiríkur Smári Sigurðarson**, University of Iceland, Iceland
4:00pm - 5:40pm Aliise Moora Auditorium	**Session SP 03: From Knowledge Graphs to Literary Networks** Location: Aliise Moora Auditorium Session Chair: **Mats Fridlund**, University of Gothenburg, Sweden, Sweden
4:00pm - 5:40pm Ilmari Manninen Auditorium	**Session SP 04: AI, Education and Language** Location: Ilmari Manninen Auditorium Session Chair: **Liisi Laineste**, Estonian Literary Museum, Estonia
5:40pm - 5:45pm	**Buses to reception**
6:00pm - 7:00pm University of Tartu Museum	**Tours at University of Tartu Museum** Location: University of Tartu Museum
7:00pm - 9:00pm University of Tartu Museum	**Welcome reception** Location: University of Tartu Museum
Date: Thursday, 06/Mar/2025	
9:00am - 11:00am Jakob Hurt’s Hall	**Session LP 04: Text Mining: Politics, Media and History** Location: Jakob Hurt’s Hall Session Chair: **Maciej Rapacz**, AGH University of Kraków, Poland
9:00am - 11:00am Helmi Kurrik Auditorium	**Session LP 05: Digital Insights in Cultural Research** Location: Helmi Kurrik Auditorium
9:00am - 11:00am	

Aliise Moora Auditorium	**Session LP 06: AI, Visualisation and Language** Location: Aliise Moora Auditorium Session Chair: **C. Annemieke Romein**, READ-COOP SCE, Austria
11:00am - 11:30am	**Coffee Break**
11:30am - 1:00pm Jakob Hurt's Hall	**Plenary Session 2: Andrea Kocsis: Can Digital Humanities Rewrite Concepts from Non-digital Heritage Studies?** Location: Jakob Hurt's Hall Session Chair: **Andres Karjus**, Tallinn University, Estonia
1:00pm - 2:00pm	**Lunch**
2:00pm - 4:00pm Aliise Moora Auditorium	**Session SP 05: Music, Film & Heritage Data** Location: Aliise Moora Auditorium Session Chair: **Haralds Matulis**, Institute of Literature, Folklore and Art of the University of Latvia (ILFA), Latvia
2:00pm - 4:00pm Helmi Kurrik Auditorium	**Session SP 06: Cultural Data: Preservation, Censorship & Interpretation** Location: Helmi Kurrik Auditorium Session Chair: **Mats Fridlund**, University of Gothenburg, Sweden, Sweden
2:00pm - 4:00pm Jakob Hurt's Hall	**Session SP 07: NLP, Language and Oral Tradition** Location: Jakob Hurt's Hall Session Chair: **Kadri Vider**, Estonian Literary Museum, Estonia
2:00pm - 4:00pm Ilmari Manninen Auditorium	**Session SP 08: OCR, Heritage & Data Accuracy** Location: Ilmari Manninen Auditorium Session Chair: **Matti La Mela**, Uppsala University, Sweden
4:00pm - 4:40pm Jakob Hurt's Hall	**Poster Session: Presentation** Location: Jakob Hurt's Hall Session Chair: **Veronika Laippala**, University of Turku, Finland
4:40pm - 5:30pm	**Poster Session 2: Posters and Demos**
5:30pm - 6:00pm	**Coffee Break**
6:00pm - 7:00pm Jakob Hurt's Hall	**DHNB: Annual General Meeting** Location: Jakob Hurt's Hall Session Chair: **Eetu Mäkelä**, University of Helsinki, Finland
6:00pm - 7:00pm	**Estonian National Museum tours**
7:30pm - 9:30pm	**Conference Dinner**
Date: Friday, 07/Mar/2025	
9:00am - 10:30am Aliise Moora Auditorium	**Panel 2: Towards Responsible DH: Practices of Critical and Caring DH** Location: Aliise Moora Auditorium Session Chair: **Mats Fridlund**, University of Gothenburg, Sweden, Sweden
9:00am - 10:30am Helmi Kurrik Auditorium	**Panel 3: Digital cultural heritage as a resource for social development** Location: Helmi Kurrik Auditorium Session Chair: **Pille Pruulmann-Vengerfeldt**, Malmö University, Sweden
10:30am - 11:00am	**Coffee Break**

11:00am - 12:30pm Jakob Hurt's Hall	**Plenary Session 3: Meelis Kull: Humans and AI: Similarities, Differences, and Why it Matters** Location: Jakob Hurt's Hall Session Chair: **Kadri Vider**, Estonian Literary Museum, Estonia
12:30pm - 1:30pm	**Lunch**
1:30pm - 3:00pm Jakob Hurt's Hall	**Panel 4: AI in the GLAM sector - Opportunities and Challenges** Location: Jakob Hurt's Hall Session Chair: **Dorna Behdadi**, Umeå University, Sweden
1:30pm - 3:00pm Helmi Kurrik Auditorium	**Session LP/SP 09: Social Media, Data and Philosophical Inquiry** Location: Helmi Kurrik Auditorium Session Chair: **Mari Väina**, Estonian Literary Museum, Estonia
1:30pm - 3:00pm Aliise Moora Auditorium	**Session LP/SP 10: Religious Geography, Translation and LLM** Location: Aliise Moora Auditorium Session Chair: **Olha Petrovych**, Estonian Literary Museum, Estonia
3:00pm - 3:30pm	**Coffee Break**
3:30pm - 5:00pm Jakob Hurt's Hall	**Final Session: Reflections and Revelations** Location: Jakob Hurt's Hall Session Chair: **Mari Väina**, Estonian Literary Museum, Estonia

KEYNOTE SESSIONS

Maciej Eder (Institute of Polish Language of the Polish Academy of Sciences)

Text Analysis Is Easy, Unless It Is Not: Reliability Issues in Measuring Textual Similarities

Text analysis investigations aimed at determining the degree of textual similarities in a collection of documents, is often associated with authorship attribution, but it can be easily generalized to address more general research questions, e.g. stylistic differentiation between genres, traces of gender, chronology, intertextuality, as well as identifying other stylometric ‘signals’. Simple as it is – at least at the first glance – the methodology used to group documents according to their similarities is at the same time based on several tacit assumptions and approximations that the users are not always aware of. The talk will revolve around a few text analysis problems, including classification, clustering, and visualization, and will focus on their limitations. A few ideas on how to improve the analysis will also be discussed.

Maciej Eder is the director of the Institute of Polish Language (Polish Academy of Sciences), chair of the Committee of Linguistics at the Polish Academy of Sciences, principal investigator of the project Computational Literary Studies Infrastructure, co-founder of the Computational Stylistics Group, and the main developer of the R package ‘Stylo’ for performing stylometric analyses. He is interested in European literature of the Renaissance and the Baroque, classical heritage in early modern literature, and quantitative approaches to style variation. These include measuring style using statistical methods, authorship attribution based on quantitative measures, as well as “distant reading” methods to analyze dozens (or hundreds) of literary works at a time.

Andrea Kocsis (University of Edinburgh)

Can digital humanities rewrite concepts from non-digital heritage studies?

With the help of a combination of distant and close reading, my paper aimed to re-evaluate why some heritage sites do not evoke hot cognition in visitors. Hot cognition is a form of affect, a direct emotional way in which we can interpret heritage experiences before or without thinking them over. Applying the term to heritage studies, David Uzzell claimed that the likelihood of the hot interpretation of a dissonant heritage site depends on the time passed between the original traumatic event and the visit. However, I argue that the exhibition's curation, the story-telling, and levels of immersion play a more critical role in the hot interpretation than the time that has passed since the atrocity. To prove so, I wanted to revise Uzzel's classic theory with new methodologies offered by digital humanities. To test my hypothesis, I have analysed 6000 TripAdvisor reviews about sites commemorating temporally distant tragedies, such as the Clifford's Tower in York, the Mary Rose Museum in Portsmouth, and the Medieval Massacre exhibition at the Swedish History Museum. While the close reading of the extant data proved fruitful and supported my hypothesis, methodologically, the research ran into a contradiction that the talk wishes to explore. I aimed to test popular computational methods in heritage affect research (sentiment analysis, lexicon-based emotion detection, topic modelling) and compare them to close reading. However, the quantitative and qualitative methods came to different conclusions. The paper investigates the reasons behind this result and points out the limitations of crowdsourced lexicon labels, lexicon-based methods deploying a categorical model of emotions, off-the-shelf codes, and the lack of control studies and methodological triangulation.

Andrea Kocsis is Chancellor's Fellow in Humanities Informatics at the University of Edinburgh. Besides being an avid writer, she has a profound interest in digital storytelling, especially in web development, facilitating the interaction between GLAM institutions and their users. She is a web developer for several research projects aiming at gamified and interactive dissemination. Her works include the digital outputs of the Unforgotten Lives exhibition at the London Metropolitan Archives. In her role as the National Librarian's Research Fellow in Digital Scholarship 2024-25 at the National Library of Scotland, she continues contributing to making the UK Web Archive's collections more accessible to wider audiences. Andrea was an Assistant Professor in History and Data Science at Northeastern University London and a Lecturer in Digital Humanities at Anglia Ruskin University before joining the University of Edinburgh.

Meelis Kull (University of Tartu)

Humans and AI: similarities, differences, and why it matters

Today's AI systems often appear strikingly human-like, operating with human signs, symbols, and texts, sometimes even making human-like errors. Yet do these systems truly understand meaning, or are we simply witnessing a sophisticated pattern-matching process that only seems meaningful? How might examining the similarities and differences between human and AI cognition help guide our choices in using these systems, and how much transparency or explainability can we realistically expect? As human identity and scholarly practices continue to evolve alongside increasingly 'thinking' technologies, how should we understand the changing role of human judgment and insight in our inquiries? The talk will explore these questions and their broader implications, encouraging critical reflection on how we shape our use of AI technologies.

Meelis Kull is a Professor of Artificial Intelligence at the University of Tartu. He is the head of the Estonian Centre of Excellence in AI. His main research topics are machine learning, artificial intelligence, and data science, with a focus on uncertainty quantification and trustworthiness.

ABSTRACTS

in alphabetical order

Anne Agersnap, Katrine F. Baunvig, Line W. Schmidt, Rie S. Eriksen, Emil W. Bønding, Thomas H. Kierkegaard, Lea W. Borcak
Aarhus University, Denmark

The Human Touch: Leveraging HITL for Quantitative Close Reading of Historical Corpora

Wednesday, 05/Mar/2025 3:00pm - 3:30pm
ID: 133 / Session LP 01: 4
Long paper (abstract) | 20-minute presentation with a 10-minute Q&A
Keywords: Quantivative Close Reading (QCR); Human-in-the-Loop (HITL); OCR; Danish Cultural History; AI-driven data analysis

1. Introduction

This paper introduces *Quantitative Close Reading* (QCR), a methodological approach that emphasizes Human-in-the-Loop (HITL) processes and manual annotation to create structured and reliable historical data from large-scale digitized corpora; it draws on established annotation procedures and adapt and apply them to the domain of digital history. QCR is particularly motivated by the poor condition of many digitized materials, particularly historical newspapers, which are frequently compromised by low-quality Optical Character Recognition (OCR) outputs. By incorporating human expertise into the annotation process, QCR corrects and refines these outputs, ensuring that the resulting dataset is accurate, nuanced, and suitable for further analysis.

This method thus bridges traditional close reading practices with computational approaches, producing high-quality, structured corpora that can be leveraged for advanced AI-driven studies. Through this synthesis of manual and automated techniques, QCR enables researchers to explore historical materials with both precision and scale, offering new insights into cultural and historical shifts.

2. HITL in Data Science: A Methodological Trend

HITL has emerged as a key trend in data science, focusing on integrating human expertise with automated systems to improve the accuracy and relevance of data processing and analysis (cf., Amershi et al. 2014). HITL methods involve human intervention at critical stages of the machine learning or data processing workflow, such as during data annotation, model training, and error correction. This approach addresses the limitations of fully automated systems, particularly when dealing with complex, nuanced data where context and human judgment are required.

In contrast to purely automated systems, HITL systems allow for iterative feedback loops, where humans validate and refine the output of machine learning models, ensuring higher accuracy and context-awareness. HITL is especially valuable for tasks involving noisy or poorly structured data, such as historical newspaper corpora, where automated tools such as OCR often fail to produce accurate results due to the quality of the original source material. One notable advantage of HITL is its ability to bridge the gap between raw data and actionable insights. Studies have shown that this hybrid approach leads to more reliable outcomes, particularly in domains such as natural language processing (NLP) and image recognition, where human intuition plays a crucial role in refining the results produced by machine algorithms (Deng et al., 2014). By involving human annotators in the manual correction and validation of OCR outputs, HITL ensures that the resulting dataset is both structured and reliable, ready for further AI-based analysis. This iterative process of human input at various stages enhances the ecological validity of the data and mitigates biases introduced by automated systems.

3. Methodology: Quantitative Close Reading and Annotation

The QCR approach involves a rigorous HITL process in which human annotators play a key role in the manual review and correction of digitized textual data.

Annotation Procedure: The QCR process begins with identifying key terms and themes that are central to the research question. In the case studies presented in the following, the terms "fællessang," "Dana," and "Lourdes" are chosen for analysis. The responsible annotator develops an annotation manual of thematic contexts to look for in the texts containing the chosen terms. The manual may consist of themes generated top-down as well as bottom-up, depending on the research question. Top-down themes may be generated from relevant theory or consist of social communities, institutions or authorities, indicative of the terms' embeddedness in society. Bottom-up themes are generated based on an initial reading of small samples from relevant periods, where the analyst infers themes that tend to occur "naturally" in the corpus, thus enhancing the ecological validity of the generated data.

Annotators search out texts containing the chosen key term, read through the texts manually and enrich the texts with metadata annotations and contextual annotations from the manual in a separate spreadsheet, thus providing quantitative data.

Validation Procedures: To ensure the accuracy of the dataset, state-of-the-art validation procedures are employed. Annotators work independently on sections of the corpus, and their annotations are validated through inter-annotator agreement metrics. This ensures that discrepancies are minimized, and any inconsistencies are reviewed and resolved by experts. Machine learning tools may be used to flag potential inconsistencies, further refining the annotations.

Dataset for AI Applications: The validated dataset produced by QCR is structured and reliable, making it suitable for computational analysis. Once refined, this dataset can be used to train AI models for further exploration, bridging the gap between manual close reading and large-scale computational analysis. This makes QCR a valuable tool for transforming compromised datasets into reliable, structured data ready for AI-driven studies.

While the QCR-approach has been applied to newspaper data in the case studies below, the method is not restricted to this data type. Whether in the form of literary texts, letters, speeches etc., digital data in a poor condition may be processed with QCR, but the annotation manual must be developed to target a specific research question and a specific dataset.

QCR requires time and personnel, making QCR a costly approach. A human annotator cannot process the same amount of data as algorithms. QCR-datasets thus tend to be considerably smaller than data for studies applying automated analyses. However, alternative solutions to solving the problem of poor data quality – such as optimizing the original OCR – are time and cost consuming as well. With QCR, preprocessing and analysis are simultaneous steps, delivering a high-quality data set.

4. Case Study: *Fællessang* (Communal Singing) in Danish Newspapers

The analysis of *fællessang* (communal singing) in Danish newspapers offers a good example of how QCR can reveal semantic shifts and stabilisations across vast corpora. *Fællessang* is a well-known participatory singing performance in Denmark. Since the 2010's, *fællessang* has gained a popular upsurge, which culminated during the COVID-19 pandemic due to media transmitted sing-alongs (Baunvig & Borcak 2023). During this upsurge, the concept seems to have undergone a singularization process (Reckwitz 2019), where predicates underpin the concept as exceptional: Singing is healthy, fun and creates social bonding (Borcak 2025). Through a QCR analysis of 4541 newspaper articles from 1850 to 2001, we find that *fællessang* historically has been categorized as a quite neutral and peripheral practice in stark contrast to nowadays underpinnings of its exceptional features (Agersnap 2025). Based on these observations, we argue that culturally potent concepts such as *fællessang* survive due to their ability of remaining constantly present in the periphery, rather in the center of attention.

5. Case Study: *Dana* in Danish Newspapers (1800–1870)

The representation of the demi-goddess *Dana* in Danish newspapers highlights the role of mythological figures in national identity formation. Using QCR, we track how *Dana* was employed in public discourse as a symbol of Denmark's mythic past. From 1800 to 1870, Danish newspapers invoked *Dana* as a symbol of Romantic nationalism, linking modern Denmark to its mythological roots. The frequent mention of *Dana* in public media served to legitimize Denmark's cultural heritage and assert national distinctiveness (Baunvig, 2021, p. 102). Through QCR, we uncover the dual role of *Dana*: while she connected Denmark to its mythic past, her image was also used to foster national unity during Denmark's Romantic reawakening (Baunvig, 2021, p. 105).

6. Case Study: *Lourdes* in Danish Newspapers (1858–1914)

The study of *Lourdes* representations in Danish newspapers illustrates how a foreign religious phenomenon was framed in public discourse over time. Using QCR to analyze 1,723 occurrences of the term 'Lourdes', we reveal key trends in Danish attitudes toward this Catholic shrine between 1858 and 1914.

Early mentions of *Lourdes* often adopted a skeptical tone, critiquing its commercial aspects and questioning the authenticity of its miracles. This reflects broader Protestant skepticism of Catholicism during the time (Baunvig, 2023, p. 59). However, by the 1890s, Danish newspapers began to express a fascination with the miraculous, driven by popular French works like Zola's *Lourdes* (1894) (Baunvig, 2023, p. 61-62). This case study demonstrates how QCR can track cultural shifts over time, highlighting evolving perceptions of faith, modernity, and commercialization in Danish public discourse.

7. QCR and *Distant Reading*: A Comparative Perspective

QCR shares with Franco Moretti's influential *Distant Reading* approach (Moretti 2013) a mutual interest in developing a method to uncover patterns across vast corpora that would be difficult to discern through traditional close reading. Though Moretti's initial conceptualization of *Distant Reading* is framed as a manual reading-through a corpus with a given focus, the emphasis in the common usage of the term is now on distant, abstract patterns, minimizing the need for close textual interpretation. QCR, in contrast, uses *Human-in-the-Loop* (HITL) procedures, where human annotators play a critical role in ensuring the accuracy and contextual relevance of the data, particularly in cases where poor OCR results compromise the dataset. This human intervention makes QCR a hybrid method that bridges large-scale analysis and traditional close reading, ensuring both accuracy and contextual depth in the analysis: QCR distinguishes itself from qualitative text analysis by generating quantitative data as the output of the manual reading procedure, instead of thick descriptions of individual texts. QCR distinguishes itself from automated distant reading by incorporating manual annotation and validation to preserve the granularity of data.

8. QCR and *Conceptual History*: A Comparative Perspective

QCR shares traits with yet another methodology - with Reinhart Koselleck's *Conceptual History*. The goal of *Conceptual History* is tracing historical shifts in meaning and focuses on key socio-political concepts, using a highly interpretive, hermeneutic approach to understand the evolution of these ideas within their historical context. It relies on close reading of canonical texts (e.g., encyclopedias) and emphasizes the philosophical depth behind these concepts. QCR, by contrast, applies a hybrid methodology that combines computational tools with human annotation to analyze large-scale digitized corpora. While both approaches study semantic changes, QCR quantifies shifts across broader textual landscapes, such as newspapers or popular media, focusing on more specific terms and phrases. In essence, while both QCR and *Conceptual History* aim to explore semantic shifts over time, QCR is more computationally driven and broader in its scope, whereas Koselleck's approach remains philosophical and contextually grounded.

9. Conclusion

QCR represents an advancement in digital humanities by offering a hybrid approach that bridges traditional close reading with computational methods. By leveraging HITL processes, QCR ensures that large-scale digitized corpora—often compromised by poor Optical Character Recognition (OCR)—are transformed into structured, reliable datasets ready for further analysis. This method not only addresses the challenges posed by noisy and inconsistent data but also preserves the contextual depth that automated systems might overlook. The case studies – of 'Fællessang', 'Dana', and 'Lourdes' – demonstrate how QCR can uncover cultural and semantic shifts in historical contexts, transforming compromised sources into valuable, interpretable datasets. Unlike Distant Reading, which emphasizes distant, abstract analysis, QCR incorporates human expertise at critical stages, allowing for both large-scale and nuanced analysis; Unlike *Conceptual History*, which focuses on deep, interpretive analysis of key socio-political concepts, QCR combines human expertise with computational tools to provide both large-scale pattern recognition and detailed, context-aware interpretation across broader textual datasets. QCR points to the fact that methods must align with data quality. QCR is a favorable approach, when the data texture is not immediately fit for computationally driven analyses. Through the synergy of manual annotation and AI-ready datasets, QCR offers researchers a comprehensive approach to exploring historical texts with precision and scale, ensuring both the integrity of the data and the richness of interpretative insights.

Bibliography

Agersnap, A. 2025. “Fællessang – Et selvfølgeligt eller et særegent ritual?”, Når vi synger – en antologi om sangens betydning, Sangens hus, 145-157.

Amershi, S., Cakmak, M., Knox, W. B., & Kulesza, T. (2014). “Power to the People: The Role of Humans in Interactive Machine Learning.” AI Magazine, 35(4), 105-120.

Baunvig, K. 2020. “Forestillede fællesskabers virtuelle sangritualer”. Sang, Vol. 1.

Baunvig, K. 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, vol. 70, 43-69.

Baunvig, K. 2021, “Fictional Realities of Modernity: The Fantastic Life of Demi-Goddess Dana in the Emerging Nation State of Denmark”, Mythology and Nation Building: N.F.S. Grundtvig and His European Contemporaries. Aarhus University Press, pp. 97-134.

Borčak, L. & Baunvig, K. 2023. ”Sang og syngning i skolen: Dansk skolesang mellem håndgribelig og uhåndgribelig kulturarv”. Tidsskriftet SANG, 4: pp. 1-26.

Borčak, L. 2025. ” ”Godhedsdiskurs” og eksklusionsmekanismer i dansk fællessangskultur”, Når vi synger – en antologi om sangens betydning, Sangens hus, 159-170.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2014). ImageNet: A large-scale hierarchical image database. In CVPR09.

Koselleck, R. 2004. Futures Past: On the Semantics of Historical Time. Translated by Keith Tribe, Columbia University Press, 2004.
Moretti, F. 2013. Distant Reading. Verso, London.

Zola, É. 1894. Lourdes.

Manex Agirrezabal Zabaleta
University of Copenhagen, Denmark

ChronoScansion: A Neural Model for the Scansion of English Poetry

ID: 121 / Poster Session 2: 30
Short paper (full-text) | 15-minute presentation with a 5-minute Q&A
Keywords: scansion, poetry, rhythm, neural network, natural language processing

In this paper we present an automatic scansion model for English poetry based on neural networks. This method is based on widely employed sequence-to-sequence models (BiLSTM and CRF), which have been proven optimal for the task. After the model is presented, we perform an experiment where the rhythm of two well known authors is checked, namely John Milton and Henry Wadsworth Longfellow. The model is made available for the research community.

Tuuli Ahlholm, Olli Nordling
The Finnish Postal Museum, Finland

From Collections to Connections: Engaging Communities via Digital Heritage in Tampere Museums

ID: 298 / WS02: 4
Innovations and New Interactions through Digital Cultural Heritage in GLAM Sector
Keywords: Museum collections, digital heritage, community engagement, born-digital, outreach

This presentation showcases past and future practical initiatives in the Finnish Postal Museum, Tampere, where both digitised and born-digital collections are employed in co-operative projects with local communities. We aim to demonstrate how even smaller GLAM organisations can successfully employ innovative digital approaches to foster deeper, more inclusive and more creative connections to heritage.

In early 2020's, the Finnish Postal Museum and Tampere Historical Museums started collaborating towards a joint application for the national status of responsibility for communication, games, post and digital life. In preparation, we launched big and small initiatives to experiment with methodologies and ways of engaging communities with digital heritage. In the 2021-2022 Erasmus+ project DREAM (Digital Reality and Educational Activities in Museums), we developed with European colleagues an experimental methodology for combining pedagogical tools with digital artefacts presented via AR. In the 2022 project *Esirippu auki!* ("Raise the curtain!") digitised artefacts were provided to communities as material for artistic expression. As a result, local groups representing intellectually disabled, senior citizens, and music students created three unique performances - all drawing inspiration from the museum's collections. Smaller projects have included outreach work on Discord servers, where we engage teenagers and young adults to document the diversities of their daily communications.

Each project produced both expected and unexpected results, and respective successes and failures. Reflecting back critically as museum professionals, we share our learning experiences: how can digital heritage be used to build up relations with different audiences and make museums - beyond just their collections - more approachable to everyone? How can we build trust with communities and encourage them to become active partners in valuing and preserving their own cultural heritage?

These questions have become all the more pressing for our institution, when in 2024 we were finally granted the status of responsibility for digital life by the Ministry of Education and Culture. Increasingly large parts of our society - from work to arts, from social life to entertainment - exist or leave traces solely in digital form. Due to the legal and technical challenges and ephemeral nature that characterise born-digital artefacts, collecting and documenting this heritage becomes an impossible task for GLAM institutions without the active co-operation of communities and creators. We share our plans for future projects, where we apply our learned lessons in order to foster meaningful collaboration and build up pride for our common born-digital heritage also.

Annastiina Ahola[1], **Eero Hyvönen**[1,2], **Heikki Rantala**[1], **Rafael Leal**[1], **Anne Kauppala**[3]
[1]Aalto University, Department of Computer Science; [2]University of Helsinki, Helsinki Centre for Digital Humanities (HELDIG); [3]University of the Arts, Sibelius Academy

Digital Humanities on the Semantic Web for visual, literary, and performing arts: Demonstrating ArtSampo, BookSampo, and OperaSampo Linked Open Data services and semantic portals

ID: 193 / **Poster Session 2: 3**
Poster and demo (abstract) with accompanying a 1-minute lightning talk
Keywords: linked data, semantic web, portals, cultural heritage

The arts can be divided into several different categories, including traditionally *visual arts*, *literary arts* and *performing arts* (Britannica 2024). This demonstration paper presents three Linked Open Data services and semantic portals for searching, exploring, and analyzing data related to these fields, and shows how they can be used to aid Digital Humanities research.

ArtSampo[1] (Ahola, Rantala, and Hyvönen 2025) is a semantic portal dealing with Finnish fine arts based on openly available art collection metadata including 80 677 art objects and 8875 people related to them (e.g., artists) from the collections of the Finnish National Gallery[2]. The data is enriched by linking to external datasets and with additional keywords and descriptions generated using multimodal generative artificial intelligence and large language models to enhance the user's experience as well as recall for keyword-based searches (Ahola, Peura, et al. 2024) and creation of recommendation links.

BookSampo[3] with its 1.6 million annual users is the most used Sampo system. It deals with Finnish literature, both works originally written in one of the national languages of Finnish and Swedish as well as translated works published in Finland. However, its original user interface does not support making literary data-analyses. As a remedy, a new alternative user interface Booksampo 2.0[4] was created (Hyvönen, Ahola, and Ikkala 2022; Ahola and Hyvönen 2023). This interface allows studying, e.g., not only the works themselves on an abstract level, but also the evolution of Finnish literature on the publication level (Ahola, Peura, and Hyvönen 2025).

OperaSampo[5] (Ahola, Hyvönen, et al. 2024) is a portal dealing with historical opera and music theatre performances in Finland in 1830–1960 with evening-specific performance data obtained from archival sources. Instead of the more traditional focus on the compositions and composers, OperaSampo follows the gradual paradigm shift in musicology that has an increased focus on also the performers and people involved with the performances (Cook 2001). This focus allows the user to, for example, easily study the history of a performer's roles on a timeline instead of the performers just being a string value in a list of names.

The portals above are part of the Sampo series of systems[6] (Hyvönen 2022) dealing with Cultural Heritage (CH) data, and are built using the Sampo-UI framework (Ikkala et al. 2022; Rantala et al. 2023). Sampo-UI allows the easy creation of new Sampo portals on top of new or existing RDF data available in SPARQL endpoints by editing JSON configurations and SPARQL query files in a declarative fashion (Rantala et al. 2023). Sampo portals enable the users to search and filter the underlying data by using the faceted search capabilities as well as perform data analysis directly in the interface with the integrated data-analytic tools without the need for programming skills. Additionally, public datasets are available on the Linked Data Finland platform[7] for more advanced querying and reuse.

[1] ArtSampo project homepage: https://seco.cs.aalto.fi/projects/taidesampo/
[2] https://www.kansallisgalleria.fi/en
[3] BookSampo – Finnish Literature on the Semantic Web, legacy system: https://www.kirjasampo.fi/
[4] BookSampo 2.0 is available at: https://analyysi.booksampo.fi
[5] OperaSampo is available at: https://oopperasampo.fi/en/
[6] https://seco.cs.aalto.fi/applications/sampo/
[7] https://www.ldf.fi/

Bibliography

Ahola, Annastiina, and Eero Hyvönen. 2023. "Visualizing Literary Linked Data for Public Library Users in the New User Interface for BookSampo – Finnish Fiction Literature on the Semantic Web." In VOILA! 2023 Visualization and Interaction for Ontologies, Linked Data and Knowledge Graphs 2023. CEUR Workshop Proceedings, Vol. 3508, July. https://ceur-ws.org/Vol-3508/paper1.pdf.

Ahola, Annastiina, Eero Hyvönen, Heikki Rantala, and Anne Kauppala. 2024. "Historical Opera and Music Theatre Performances on the Semantic Web: OperaSampo 1830-1960." In Knowledge Graphs in the Age of Language Models and Neuro-Symbolic AI: Proceedings of the 20th International Conference on Semantic Systems, 17–19 September 2024, Amsterdam, The Netherlands, 386–402. Studies on the Semantic Web. IOS Press, September. https://doi.org/10.3233/SSW240031.

Ahola, Annastiina, Lilli Peura, Rafael Leal, Heikki Rantala, and Eero Hyvönen. 2024. "Using generative AI and LLMs to enrich art collection metadata for searching, browsing, and studying art history in Digital Humanities." In Proceedings, 2nd International Conference on Data & Digital Humanities: Generative Artificial Intelligence for Text and Multimodal Data 12th - 13th December 2024, University of Minho, Braga, Portugal. Accepted, forth-coming.

Ahola, Annastiina, Telma Peura, and Eero Hyvönen. 2025. "Using linked data for data analytic literary research: Case BookSampo—Finnish fiction literature on the semantic web." Journal of the Association for Information Science and Technology, https://doi.org/10.1002/asi.24984.

Ahola, Annastiina, Heikki Rantala, and Eero Hyvönen. 2025. "ArtSampo – Finnish Art on the Semantic Web." In The Semantic Web: ESWC 2024 Satellite Events, edited by Albert Meroño Peñuela, Oscar Corcho, Paul Groth, Elena Simperl, Valentina Tamma, Andrea Giovanni Nuzzolese, Maria Poveda-Villalón, et al., 159–163. Cham: Springer Nature Switzerland.

Britannica, The Editors of Encyclopaedia. 2024. The arts. Accessed on October 16, 2024. https://www.britannica.com/topic/the-arts.

Cook, Nicholas. 2001. “Between process and product: Music and/as performance.” Music theory online 7 (2): 1–31.

Hyvönen, Eero. 2022. “Digital Humanities on the Semantic Web: Sampo Model and Portal Series.” Semantic Web 14 (4): 729–744. https://doi.org/10.3233/SW-223034.

Hyvönen, Eero, Annastiina Ahola, and Esko Ikkala. 2022. “BookSampo Fiction Literature Knowledge Graph Revisited: Building a Faceted Search Interface with Seamlessly Integrated Data-analytic Tools.” In Theory and Practice of Digital Libraries (TDPL 2022), Accelerating Innovations Track, Padova, Italy, 506–511. Springer-Verlag. https://doi.org/10.1007/978-3-031-16802-4_54.

Ikkala, Esko, Eero Hyvönen, Heikki Rantala, and Mikko Koho. 2022. “Sampo-UI: A Full Stack JavaScript Framework for Developing Semantic Portal User Interfaces.” Semantic Web 13(1): 69–84.

Rantala, Heikki, Annastiina Ahola, Esko Ikkala, and Eero Hyvönen. 2023. “How to create easily a data analytic semantic portal on top of a SPARQL endpoint: introducing the configurable Sampo-UI framework.” In VOILA! 2023 Visualization and Interaction for Ontologies, Linked Data and Knowledge Graphs 2023. CEUR Workshop Proceedings, Vol. 3508. https://ceur-ws.org/Vol-3508/paper3.pdf.

Agnes Aljas[1], Pille Pruulmann-Vengerfeldt[2], Pille Runnel[1]
[1]Estonian National Museum, Estonia; [2]University of Malmö

Innovations and New Interactions through Digital Cultural Heritage in GLAM Sector

ID: 283
Half-day conference-themed workshop
Keywords: digital cultural heritage, innovations, co-creation, audience engagement, new forms of interaction

The workshop focuses on the application of digital cultural heritage, which has significant potential to contribute to the sustainable renewal of society. Reinterpreting cultural heritage needs to be addressed from social, digital, and green transition perspectives (Caro-Gonzalez, 2023). The UNESCO and FARO conventions prioritize human-centered heritage management and equal access to heritage, indicating that the value of heritage can be measured by what it adds to people's lives. Digitization of heritage is seen as an important tool for overcoming the dilemma between preservation and usage (Parry 2010, Cameron 2021), but today the focus is often on creating digital repositories and less on how different societal groups could use digital heritage for societal development.

Although museums are recognized as an important sector contributing significantly to economic growth, fostering innovation, and spreading knowledge, their potential societal benefit has not been fully understood or thoroughly researched (Falk 2021, 2022, Social... 2011, Carmen 2010, Packer 2010). Research has shown that cultural heritage plays a crucial role in supporting people's ability to respond to societal changes and contribute both personal and social well-being (Crossick 2016, Carnwath 2014, Falk 2022, Brown and Novak 2013; Brown and Ratzkin 2011; Walmsley and Franks 2011), as well as the local institutional ecosystem (Hudson 2015, Caro-Gonzales 2020). Here, co-creation of experiences between stakeholders (Minkiewitz 2014) has potential (Jaakkola 2015, Chaney 2012). Digital cultural heritage is an important resource for museums that should be used much more today in data-driven decisions (Pruulmann-Vengerfeldt, 2022), co-creation initiatives, citizen science, and the education sector more broadly.

We invite proposals for presentations that facilitate discussions on the applications of digital technology within the GLAM sector, seeking to showcase and analyse innovative approaches to preserving, communicating, and interpreting cultural heritage. We are especially interested in cases that leverage digital cultural heritage for co-creation activities, audience participation, and other user engagement initiatives, which would support utilizing the potential of digital cultural heritage across cultural, educational, economic, and other fields.

We welcome contributions that highlight developments enabled by digital technologies that enhance accessibility and engagement to heritage institutions, creating opportunities for new audiences and new partnerships.

By introducing practical use cases, as well as presenting critical studies about the challenges and opportunities brought by the digital turn in the heritage sector, we aim to bring together practitioners and scholars. Our goal is to foster discussions and enhance understanding of the profound and transformative impact that digital innovation can have on cultural heritage. Proposals should aim to contribute to a better understanding of how technology can not only preserve and interpret cultural heritage but also serve as a bridge to new forms of interaction and collaboration.

Information about the target audience

- Researchers and practitioners in the GLAM sector who are exploring or implementing digitalized collections, tools, and strategies to enhance their work. This includes those integrating new digital innovations into long-term institutional strategies and focusing on co-creation, accessibility, and user engagement within cultural institutions.
- Researchers who have been using digital humanities methods in heritage research and analysis
- Developers, designers, and technology providers working on digital solutions for cultural heritage applications.

Expected Outcomes

- The workshop aims to discuss new ways to interpret and utilize digital cultural heritage.
- Explore topics such as accessibility, sustainability, and audience diversity in digital cultural heritage projects.
- Provide participants with an opportunity to share innovative ideas on how to integrate digital technologies into GLAM institutions, focusing on user engagement, relevance, and collaboration.

Daniel Antal[1], Britt-Kathleen Mere[2], Anna Márta Mester[1], Kata Gábor[3], Ieva Pigozne[4], Bogáta Tímár[2], Fábio Generoso Vieira[6], Ieva Vīvere[5]

[1]Reprex B.V, the Netherlands; [2]University of Tartu, Estonia; [3]INALCO, France; [4]Institute of Latvian History, Latvia; [5]Archives of Latvian Folklore, Latvia; [6]University of Amsterdam, the Netherlands

A Finno-Ugric Data Sharing Space

ID: 252 / **Poster Session 2: 24**
Poster and demo (full-text) with accompanying a 1-minute lightning talk
Keywords: Wikibase, Dataspace, Trustworthy AI, Open knowledge

A concept and a demo of a *Finno-Ugric Data Sharing Space* as a knowledge base and a trustworthy AI application, and a replication of our *Slovak Comprehensive Music Database* created with a data sharing space in 2020-2024 . Our application follows the novel AI and data regulatory requirements and recommendations, particularly human-in-control and human-in-the-loop procedures, with a particular emphasis on community stewardship and control.

We present a small scale dataspace as an interoperable, interdisciplinary, and area-relevant explicit knowledge base and trustworthy AI that poster viewers can try out. To show the versatility of our cross-domain conceptualisation, we will connect songs about “Dreams” with potential locations, performances, choreographies and festive dress from various ethnic groups of the Baltics area.

Daniel Antal[1], Britt-Kathleen Mere[2], Anna Márta Mester[1], Kata Gábor[3], Ieva Pigozne[4], Bogáta Tímár[2], Fábio Generoso Vieira[6], Ieva Vīvere[5]

[1]Reprex B.V, Netherlands, The; [2]University of Tartu, Estonia; [3]INALCO, France; [4]Institute of Latvian History, Latvia; [5]Archives of Latvian Folklore, Latvia; [6]University of Amsterdam, the Netherlands

Finno-Ugric Data Sharing Space: Federating Open Knowledge About Contemporary and Historic Cultural Practices in the Wikibase System

ID: 256 / Session LP 05: 4
Thursday, 06/Mar/2025 10:30am - 11:00am

Long paper (abstract) | 20-minute presentation with a 10-minute Q&A
Keywords: Dataspace, Wikibase, Open knowledge, Open data, Finno-Ugric

In our paper we would like to present a demonstration of a trustworty artificial intelligence application filled with actual datasets about contemporary and historic Finno-Ugric cultural practices centered around music. Our paper intersects with the topics of *Integrating traditional humanities and computation* and *Artificial Intelligence*, but perhaps best answers the *Coming down the 'Ivory Tower'*. We want to demonstrate is possible to federate a well-designed, and existing data (sharing) space utilising the connection of the Wikibase open-source knowledge management and the R open-source language with novel disciplinary, institutional, geographical, or culturally and linguistically broader data (and knowledge.)

We would like to answer the following questions:

- How is it possible to extend GLAM data interoperability programs to the private sector, particularly in areas like music, audiovisual or fashion, where the private sector holds more data and digital assets than GLAM institutions?
- What measures can be taken to avoid the data biases described in intersectional data feminist critique that lead to an underrepresentation of women and ethnic minorities in databases?
- To what extent can we connect strictly curated GLAM knowledge with authority control with subcultural and ethnic minority knowledge that is often tacit or described at best with folksonomies?
- To avoid the bad experience with the Liv and Mari Wikipedia incubators, is it possible to improve the community stewardship, data curation and community control practices of open knowledge management?
- Generally, how can a trustworthy AI application, i.e., an intelligent knowledge base, comply best with the recent Data Governance Act and the AI Act of the European Union? What lessons are learned about data protection when joining public and private-sector data?

1.1. Methodology

Utilising our experience and own open-source, peer-reviewed scientific software code, we will create datasets that can be turned into knowledge (semantically valid knowledge statements with an actual/false value). We add these datasets to a semantic knowledge base with our own developed extensions of the Wikibase open knowledge management system. This way we create an open graph database that offers an interoperable schema with a potential to connect various organisations' or research institutions'databases.) We will showcase our dataspace developed in Slovakia with federated new datasets on contemporary Finno-Ugric popular music, historical folk music of the Baltic area, dress and dance history of the Baltic Area, authority control data on named entities and terms, and generally useful metadata and compare them with our far more comprehensive results in Slovakia.

We want to create tools for an improved application of the European Interoperability Framework that not only connects public GLAM and open science data resources but also privately-held data. This approach requires significant contributions in data governance (the legal and organisational aspects of interoperability) and some improvements in the semantic and technology layers.

The *Wikibase Data Model* and the Wikibase open knowledge management system have been successfully used in many digital humanities interest projects to build an interdisciplinary consensus on conceptualising the heritage of interest. From a data science point of view, our main effort is connecting the Wikibase open knowledge management system with the R statistical environment and creating new extensions of both systems (Daniel Antal 2022, 2024); in our paper, we would like to recall the functional requirement setting with software- and ontology design patterns and the released (and peer-reviewed code.)

Our overarching methodology for connecting computer and data sciences with digital humanities and copyright law is appropriate conceptualisation and ontologies encoded into a collaborative semantic system that is usable by both digital humanities researchers and private data owners.

1.2. Public-private interoperability

With the European Commission's support in the Open Music Europe project (Open Music Europe 2023), we have been implementing a data-sharing space in Slovakia for GLAM organisations and private parties [Commission et al. (2020); (Daniel Antal 2020). The European Union's objective is to improve the interoperability of open science and statistics towards the business sector, particularly important with humanities research materials of music, film and fashion. In these areas, copyrights or neighbouring rights often apply, which creates legal and organisational barriers to digital humanities research. Private parties manage the global copyright and neighbouring rights identification and registration system, the intellectual rights of such humanities-interest material private property, and its authority file systems. Private parties also own a magnitude larger amount of digital and data assets in music, film, and fashion.

Bringing down legal and organisational barriers is an important first step, but if we cannot go further on this road with a semantic and technological alignment, then data will still not be interoperable. Our aim is not to optimise the performance of our ontologies and knowledge basis regarding inferential capacity, or to compete with large, basic-research type ontology design work like that of Polifonia (Berardinis et al. 2023), but to design software and ontology patterns (Blomqvist, Hammar, and Presutti 2016) in R and RDF (data/metadata languages) that can break through the barriers in a cost-effective way without perfection. We will show this approach with connecting knowledge about music, estive dresses, dance choreography, and lyrics.

In the 2020s, statistician increasing abandon questionnaire-based surveying (which is costly and often inaccurate), and tap into the data systems of electricity companies, digital platforms, and other private administrative data sources to find more accurate and up-to-date information (ESS 2022). We would like to show that these novel approaches to data collection and consolidation have a high reuse potential in digital humanities. Statisticians and socio-economic researchers want to avoid systematic biases when they collect data; for example, they cannot utilise a system that provides less accurate data on women musicians than men. While in humanities research, the aim of curating novel collections for analysis does not aim at statistical representativeness, the statistical methodology of designing collections that avoid biases appears to be particularly useful for researching areas that had been overlooked by an intersection of past discrimination, for example, because historically, less data has been collected about women, members of small ethnic groups, and especially stateless minorities. (D'Ignazio and Klein 2020)

1.3. Authority control

The incomplete, "as-needed" way of conceptualisation, in line with the spirit of dataspaces, provides us encouraging results in breaking down not only institutional data silos but disciplinary ones. While we developed our system with a music focus, we successfully applied it to textile research (Pigozne and Antal 2024). We want to demonstrate our ability to cross-domain conceptualise, and we will connect songs around the rather fuzzy concept of *Dreams* with potential locations, performances, choreographies, and festive dresses from various Baltic nations.

In digital humanities, authority control for named entities (people and things described in thesauri) is a key requirement for joining data collections across GLAM institutions. The same can be said when we extend access to private data sources; however, authority control becomes far more complicated. In our paper, we want to publish the first case in which we have legally and operationally connected a national copyright registry (Slovak Performing and Mechanical Rights Society) parallel to the authority control of a national library.

When we approach smaller countries or mainly stateless ethnic minorities, authority control often cannot rely on explicit (already interoperable and semantically coded) authoritative information. We must be able to conceptualise and describe the tacit knowledge of communities appropriately, something we will show with a recent collection of Finno-Ugric minority popular music.

The independently produced music of Baltic subcultures, not to mention that Finno-Ugric popular music often recorded in the Russian Federation is missing from copyright registers. Their authors are not present in library systems, and their music is described on a folksonomy level at best. We want to tool and empower curators of less formally described knowledge about Finno-Ugric cultural practices to have a toolkit and support to connect knowledge about these communities, subcultural groups, and their music, dance, fashion or other cultural practices to a genuinely interoperable knowledge base. We would also like to show how the results of a groundbreaking (explicit knowledge creation) work, the results of the *Documenting and Mapping Livonian Place Names and Creating an Official Place Name Register (2020–2022) project* (Ernštreits 2020) can be utilised in this regard.

Wikibase has often been used for authority control (Bianchini, Bargioni, and Pellizzari di San Girolamo 2021; Fagerving 2023) with excellent results. However, we are also informed about the highly problematic incubated Liv and Mari Wikipedias, which researchers and minority communities generally see as bad examples of language corpus pollution. In our paper, we want to show how improved data governance in the co-curation of collaborative open graph databases can connect already authoritatively described data and metadata to the scattered knowledge of communities or novel research that had not previously relied on artificial intelligence or semantic technologies.

1.4. Inclusive Data Governance and Trustworthy AI

A dataspace is an emerging approach to data management which recognises that in large-scale integration scenarios involving many partners, it would be prohibitively expensive and time-consuming to obtain an upfront unifying schema across all sources or to come to a legal agreement on the terms of using or exchanging the data. It is an intelligent application that allows a near-instantaneous exchange, processing, sharing and provision of data on an "as-needed" or "as-permitted" basis while retaining complete control of each data holder over the conditions (e.g., who, when, and under what condition) of access to their data (Curry 2020; EBU and Gaia-X 2022; Nagel and Lycklama 2021).

Exchanging data on an "as-needed" basis requires trust among organisations that have never exchanged data before. In Slovakia, we kept involving all key stakeholders. Eventually, we created a high-level Memorandum of Understanding with key stakeholders (Ministerstvo kultúry SR and Open Music Europe 2023), which was later extended with further written protocols among the authority control bodies of the national library and the copyright registration agency on one hand and with a pioneering engagement of the Wikipedian community via a special Wikipedian in Residence program (Dániel Antal, Grochal, and Varvantakis 2024). With much smaller datasets and less at stake, we would like to bring this approach a step closer to small subcultural or ethnic minority communities in the Baltic region by means of data federation and opening up or ealier geographically Central Europe-centered system.

Trust can only be maintained with a robust, often codified, well-understood governance model that applies strict normative guidelines on how data sharing can be organised and executed. An important inspiration to our work is the creation of the Luxembourg Shared Authority File with Wikibase. While remaining in the realms of the public GLAM sector, the project already faced challenges with applying the European General Data Protection Rules (GDPR) in GLAM (Pfeiffer and Gayo 2021). In our work, we had to bring data governance to a next level, because private legal entities must apply GDPR in a more restrictive manner.

In late 2024, the decade-long foundational work about autonomous systems' ethical and human-rights aspects (European Union Agency for Fundamental Rights 2020; Commission, Directorate-General for Communications Networks, and Technology 2019) is kicking in as binding regulation in the form of the Data Governance and AI Act of the EU.

The top priority in building a trustworthy AI system is to provide human control or at least keep humans in the loop. In this regard, we would like to apply novel language technology extensions of the Wikibase system, which allow the reverse translation of semantic statements (machine-coded knowledge) to natural human language, including, to some extent, minority languages. We will do this with adding to the Wikibase system that coordinates or data sharing space a linguistically well-configured *Lexeme* extension that provides improved modelling for lexical entities such as words and phrases, and to improve re-use and mappings to other vocabularies. We would like to create a toolset that allows the revision of the knowledge encoded in our shared open graph database in many human languages and in terms that it is understandable and transparent to formal data owners or community data stewards of contemporary popular music or ethnic minority heritage; then negotiate data exchange rules that will eventually lead to a consensual governance of the shared, open, linked database on Finno-Ugric material and immaterial cultural heritage.

Bibliography

Antal, Daniel. 2020. “Feasibility Study on Promoting Slovak Music in Slovakia & Abroad.” https://doi.org/10.5281/zenodo.6427514.

Antal, Daniel. 2022. “dataset: Create Data Frames that are Easier to Exchange and Reuse.” The Comprehensive R Archive Network. https://doi.org/10.5281/zenodo.7440192.

Antal, Daniel. 2024. “Making Datasets Truly Interoperable and Reusable in R. The specific case of working with the Wikibase Data Model.” Zenodo. https://doi.org/10.5281/zenodo.13972087.

Antal, Dániel, Michal Grochal, and Christos Varvantakis. 2024. “Building a Music Data Sharing Space with Wikibase.” Zenodo. https://doi.org/10.5281/zenodo.8046977.

Berardinis, Jacopo de, Valentina Anita Carriero, Nitisha Jain, Nicolas Lazzari, Albert Meroño-Peñuela, Andrea Poltronieri, and Valentina Presutti. 2023. “The Polifonia Ontology Network: Building a Semantic Backbone for Musical Heritage.” In The Semantic Web – ISWC 2023, edited by Terry R. Payne, Valentina Presutti, Guilin Qi, María Poveda-Villalón, Giorgos Stoilos, Laura Hollink, Zoi Kaoudi, Gong Cheng, and Juanzi Li, 302–22. Lecture Notes in Computer Science. Cham: Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-47243-5_17.

Bianchini, Carlo, Stefano Bargioni, and Camillo Carlo Pellizzari di San Girolamo. 2021. “Beyond VIAF Wikidata as a Complementary Tool for Authority Control in Libraries.” Information Technology and Libraries 40 (2). https://doi.org/10.6017/ital.v40i2.12959.

Blomqvist, Eva, Karl Hammar, and Valentina Presutti. 2016. “Engineering Ontologies with Patterns – the eXtreme Design Methodology.” In Ontology Engineering with Ontology Design Patterns, 23–50. IOS Press. https://doi.org/10.3233/978-1-61499-676-7-23.

Commission, European, Content Directorate-General for Communications Networks, and Technology. 2019. Ethics Guidelines for Trustworthy AI. Publications Office of the European Union. https://doi.org/doi/10.2759/346720.

Commission, European, Sport Directorate-General for Education Youth, Culture, M Clarke, P Vroonhof, J Snijders, A Le Gall, et al. 2020. Feasibility Study for the Establishment of a European Music Observatory : Final Report. Publications Office. https://doi.org/doi/10.2766/9691.

Curry, Edward. 2020. “Dataspaces: Fundamentals, Principles, and Techniques.” In Real-Time Linked Dataspaces: Enabling Data Ecosystems for Intelligent Systems, 45–62. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-29665-0_3.

D’Ignazio, Catherine, and Lauren F. Klein. 2020. Data Feminism. Strong Ideas. Cambridge, MA, USA: MIT Press. https://data-feminism.mitpress.mit.edu/.

EBU, and Gaia-X. 2022. “Dataspace for Cultural and Creative Industries. Position Paper. v.2.0.” Gaia-X. https://gaia-x.eu/wp-content/uploads/2022/10/EBU_position-paper_Media-Data-Space.pdf.

Ernštreits, Valts. 2020. “Livonian Place Names: Documentation, Problems, and Opportunities.” Eesti Ja Soome-Ugri Keeleteaduse Ajakiri. Journal of Estonian and Finno-Ugric Linguistics 11 (1): 213–33. https://doi.org/10.12697/jeful.2020.11.1.09.

ESS. 2022. Privately Held Data Communication Toolkit. 2022nd ed. Manuals and guidelines. Luxembourg: Publications Office of the European Union.

European Union Agency for Fundamental Rights. 2020. Getting the Future Right. Artificial Intelligence and Fundamental Rights. Luxembourg: Publications Office of the European Union. https://fra.europa.eu/sites/default/files/fra_uploads/fra-2020-artificial-intelligence_en.pdf.

Fagerving, Alicia. 2023. “Wikidata for Authority Control: Sharing Museum Knowledge with the World.” Digital Humanities in the Nordic and Baltic Countries Publications 5 (1): 222–39. https://doi.org/10.5617/dhnbpub.10665.

Ministerstvo kultúry SR, and Open Music Europe. 2023. “Memorandum o porozumení o využití výsledkov analýz otvorených politík v kontexte slovenského kultúrneho a kreatívneho priemyslu a sektorových verejných politík v spolupráci s konzorciom pre výskum a inovácie s názvom OpenMuse. [Memorandum of Understanding on utilizing the Open Policy Analysis results of the OpenMuse Research and Innovation Consortium in the context of Slovak cultural and creative industries and sectors’ public policies].” https://www.crz.gov.sk/zmluva/7645338/.

Nagel, Lars, and Douwe Lycklama, eds. 2021. “Design Principles for Data Spaces. Position Paper. Version 1.0.” Open DEI. https://doi.org/10.5281/zenodo.5244997.

Open Music Europe. 2023. “Open Music Europe (OpenMusE) – An Open, Scalable, Data-to-Policy Pipeline for European Music Ecosystems.” https://doi.org/10.3030/101095295.

Pfeiffer, Michelle, and Jose Emilio Labra Gayo. 2021. “Representing CIDOC-CRM in Wikibase for the Luxembourg Shared Authority File.” https://swib.org/swib21/slides/05-03-gayo.pdf.

Pigozne, Ieva, and Dániel Antal. 2024. “Linked Open Datasets on Garments from the Latgale Region.” Zenodo. https://doi.org/10.5281/zenodo.13971707.

Johanna Arnesson, Evelina Liliequist, Coppélie Cocq
Umeå University, Sweden

Collecting memories of the early internet

ID: 164 / Session LP 02: 2
Wednesday, 05/Mar/2025 2:00pm - 2:30pm
Long paper (abstract) | 20-minute presentation with a 10-minute Q&A
Keywords: internet histories; digital practices; qualitative methods

Today, it is perhaps difficult to imagine a world where the internet is not an integral part of both everyday life and various societal functions in Sweden. Especially for those who were born after the turn of the millennium and thus grew up in a time when digital technology is taken for granted. But once upon a time, the internet was new and exciting. Conversations about personal memories of the early internet sparked our interest in knowing more about what people actually did on the internet when it was new. What do every day users remember from that time and what was the impact of this new technology?

From March to August 2024, the questionnaire Mötet med internet (Encounter with the internet) was available on the website of the Swedish Institute for Language and Folklore (ISOF). The aim was to collect stories about early internet use in Sweden, focusing on everyday practices and users' memories of, and reflections on, the internet during (predominantly) the 1990s and early 2000s. We asked questions such as "Do you remember the first time you heard about the internet? When and in what context was it, and how old were you then? What did you do on the internet and how did you access it?". It was answered by 225 people, of whom 130 also expressed interest in being interviewed on the topic. The questionnaire forms the basis of a pilot study conducted by our interdisciplinary research group in Digital Humanities at Humlab, Umeå University.

In this paper, we will first present preliminary results from the questionnaire, based on an initial thematic overview of the responses. Second, we discuss methodological challenges associated with these types of sources, as well as what kind of media histories and ethnographic knowledge they can contribute to.

Research on the history of the internet often focuses on the individuals, infrastructures, stakeholders and funders that were influential in the development of networked computer communication and what came to be the internet. In contrast, we want to document and explore these developments from a user perspective, focusing on the practices, experiences and perceptions of 'ordinary people' on the early internet.

In recent research, it is argued that the grand narratives of the internet are often limited and reduced to singular stories about ARPANET and the Silicon Valley, overlooking the myriads of user-developments achieved through early experimental practices such as hobby radio, BBSs, file-sharing etcetera (Driscoll 2022: 194). In this project we therefore want to give voice to users' everyday experiences in their early explorations of the internet in a Swedish context. We are specifically interested in stories from those who did not have a prior interest in and experience of computer networks and digital communication, but for whom the internet still had an impact on their lives since they experienced the introduction of the commercial internet and the World Wide Web. The choice of geographical context is partly motivated by practical reasons, since it is the context in which we are active as researchers. However, Sweden is also a particularly interesting case since it is a country that, through political initiatives and economic priorities, was an early adopter of the internet among its citizens. For example, in 1994, the government started the first "IT commission", with the explicit mission to initiate and facilitate digital development in the country (SOU 1994:118). Among other things, almost 1 billion SEK was invested in information technology for education and schools. The same year, Algonet, the first commercial internet service provider aimed at ordinary Swedes, was launched.

The analysis of the responses from the questionnaire enables us to discern some patterns in the material. The answers vary from short descriptions to longer stories with many details. In the somewhat longer answers, a certain nostalgia can be detected, often in combination with a predominantly positive description of early encounters with online environments. At the same time, some respondents recall how boring the internet was – they did not know what to look for, how to find it (there were no search functions), or what to do with it. Most respondents have come into contact with the internet either through their work or through social relationships (for example, a friend interested in technology, at their parents' job, or through an uncle with a computer at home). The answers also raise new questions, such as what exactly counts as 'the internet'. Some people came into contact with technologies that network computers as early as the 1970s. For most, however, the popularisation of the internet with the introduction of the World Wide Web (WWW) in the early 1990s seems to be the decisive factor.

Specific practices such as searching for information, chatting, forum discussions, and work-related use are mentioned in the responses. The stories show that for some the internet was gradually integrated into their everyday and working life, while for others it was perceived as a revolution. There are also recurring descriptions of financial and technical constraints – the internet could only be used after the magic hour of 6pm, when it was cheaper, and downloading a very blurry image could take several hours. Other technical aspects also emerge – memories of the scraping and beeping sound of a modem connecting to the internet, for example, are mentioned in several responses. Thus, the material also reflects sensory and embodied memories. Collective use is another thing that stands out in the materials. Especially for those who were young at the time, the internet was something used together with siblings, friends, or classmates, often on a shared computer located in a 'public' environment, at home or at school. This is quite a contrast to today's individualised use, where most people have 'their own' internet, both in terms of technology and content – an internet that also is mobile and wireless rather than bound to a specific location and connection.

Methodologically, this material offers an opportunity to know more about early internet use in Sweden, and to design a media ethnography based on more thorough interviews and historical sources. In her study of young people in Canadas experiences of going online, MacKinnon (2022b) argue that "web archives alone are insufficient in reconstructing the experience of going online" (p. 4) and suggests that a patchwork of methods and materials, providing different sources and perspectives, is needed. In similar ways, our pilot study will draw on both oral and written sources, as well as internet archival samples. Through participants' written and oral histories about the internet in retrospect, we will explore people's meaning-making about the development and impact of the internet. The questionnaire and subsequent interviews will be combined with historical sources including Swedish Government Official Reports such as IT commission (SOU 1994:18) and media content from the time (news media, magazines, TV programmes etcetera).

Paßmann (2021) and Paßmann & Gersen (2024: 214), advocates for “elicitation interviews through archived web materials as part of a qualitative framework for digital methods”. We also see methodological potentials of elicitation through web archives for our research. Firstly, it is a way to help people remember, and secondly, to compare what people recollect versus what the digital environments they talk about actually looked like. The point of this is not to verify peoples’ memories, but rather to examine what is remembered, how it is remembered, and why. As part of this endeavour, we will conduct ‘archive promenades’, where the respondent, together with the researcher, looks back into their own internet history with the help of the Internet Archive (Mackinnon 2022a: 356-357).

However, the approach also comes with some critical issues. For example, there is a certain bias in the material, both because of the particular interest of those who responded to the questionnaire and because of the inherent challenges of using people’s memories and stories as historical sources. The questionnaire responses (and subsequent interviews) capture how people remember and describe their past experiences today. Previous experiences of archive promenades show that this approach requires that respondents have a prior interest in their own digital history and have ideas about how to access it, such as usernames, webpage addresses, or what sites to search for (MacKinnon 2022b:45). A challenge, then, is to keep the focus on what the internet was like then, without too much reflection on what we know about what has happened after the time period, and the digital landscape today (Brügger & Goggin, 2022:2). At the same time, participants’ reflections linking the past to the present are analytically interesting, as this can shed light on how past experiences have shaped their everyday lives, relationships, and working lives, but also how ‘ordinary’ people’s use has shaped the internet.

Furthermore, there is also a risk that the stories received via the questionnaire and in future interviews will be characterised by an over-romanticisation of the past. This may mean that potential problematic experiences among the users, such as exclusion, marginalisation, discrimination, conflicts or similar, are left out of the narratives. Simultaneously, the stories from those who say that the internet did not have a major impact in their lives, and who mainly mention mundane, everyday use in their responses, should not be overlooked.

The risk of bias is also linked to an issue of representation – who responded to the questionnaire and how will we reach potential interview participants? Those who chose to respond to the questionnaire are a defined group with specific experiences related to the dissemination and outreach of the questionnaire. The questionnaire was disseminated through ISOF:s distribution channels, and our own private and academic networks. If it had been shared and distributed by, for example, the National Museum of Science and Technology or the (online) Internet Museum in Sweden, more responses might have been submitted from people with rich personal stories about networked communication and internet practices. However, that might have skewed the material even more towards a tech-centric narrative, and less towards the everyday user experiences.

Bibliography

Brügger, N., & Goggin, G. (2022). Oral Histories of the Internet and the Web: An Introduction. In Oral Histories of the Internet and the Web (pp. 1-8). Routledge.

Driscoll, K. (2022). The modem world: A prehistory of social media. Yale University Press.

Mackinnon, K. (2022a). Critical care for the early web: ethical digital methods for archived youth data. Journal of Information, Communication and Ethics in Society, 20(3), 349-361.

Mackinnon, K. (2022b). Databound: Histories of Growing Up on the World Wide Web (Doctoral dissertation, University of Toronto (Canada)).

Paßmann, J. & Gerzen, L. (2024) “Follow the updates! Reconstructing past practices with web archive data”, Internet Histories, 8:3, 213-228, DOI:10.1080/24701475.2024.2310405

Paßmann, J. (2021). Medien-theoretisches Sampling Digital Methods als Teil qualitativer Methoden. Zeitschrift für Medienwissenschaft, 13(25-2), 128-140. https://doi.org/10.14361/zfmw-2021-130213

SOU 1994:18 “Informationsteknologin - Vingar åt människans förmåga. Betänkande av IT-kommissionen"

Merilin Aruvee, Kais Allkivi, Andres Karjus, Krister Kruusmaa, Katarin Leppik, Silvia Maine, Taavi Kamarik, Harli Kodasma

Tallinn University, Estonia

Automated writing evaluation of Estonian as a first language

ID: 240 / Poster Session 2: 9

Poster and demo (abstract) with accompanying a 1-minute lightning talk

Keywords: first language teaching, writing instruction, natural language processing, automated writing evaluation

The transition to Estonian e-exams in 2026 has highlighted the necessity to create efficient and versatile automated writing evaluation tools that would support teachers in giving feedback and assessing writing tasks. A number of such tools have been developed for English teachers and learners (i.e., Criterion, MyAccess, Write & Improve) but Estonian language teachers lack similar opportunities. On the other hand, similar work has so far focused on measuring second language, rather than first language proficiency.

Our 2-year project draws on the need to develop computer-assisted language learning and writing assistant software for Estonian, which could facilitate classroom and independent learning, and help teachers check written assignments and support exam assessment.

The objectives are: 1) to identify optimal solutions for evaluating argumentative texts of the 9th and 12th grade e-exams, using language technology; 2) to develop a web application, which offers writing recommendations, predicts the grade of the text and provides feedback based on the exam evaluation rubric.

Project team works closely with Estonian Education and Youth Board who has provided access to e-exam texts from the year 2024 (currently, the data consists of 795 argumentative texts from 9th graders and 764 texts from 12th graders). We analyze texts by using corpus analysis, supervised machine learning, and large language models, evaluating vocabulary, grammar, correctness as well as text content and structure.

This poster introduces the preliminary results of the research.

Yan Asadchy, Maximilian Schich
Tallinn University, Estonia

Descriptions of men are shorter than of women: prompt analysis of gender portrayals in Stable Diffusion

ID: 180 / Poster Session 2: 22
Poster and demo (abstract) with accompanying a 1-minute lightning talk
Keywords: Generative AI, AI Art, Gender Stereotypes, Representations

Generative AI for image creation is becoming a staple in the toolkit of digital artists, visual designers, and the general public who want to be represented through gen AI. Social media users have many tools to shape their visual representation: image editing tools, filters, face masks, face swaps, avatars, and AI-generated images. The importance of the right profile image can not be understated: it is crucial for creating the right first impression, sustains trust, and enables communication. Correct representation of individuals, groups of people, and collectives helps to foster inclusivity, understanding, and respect in society, ensuring that diverse perspectives are acknowledged and valued. While previous research revealed the biases in large image datasets such as ImageNet and inherited biases in the AI systems trained on it, in this work, we look at the biases and stereotypes as they emerge from textual prompts used for generating images on Discord with StableDiffusion model. We analyze over 2.5 million prompts depicting men and women and use statistical methods to uncover how prompts describing men and women are constructed and what words constitute the portrayals of respective genders. Our findings suggest uniform practice of prompting regarding word length; however, the optimal men's descriptions are shorter than those of women. When it comes to word and topic analysis, our findings suggest the existence of classic stereotypes in which men are described using dominant qualities such as "strong" and "rugged". In contrast, women are represented with concepts related to body and submission: "beautiful", "pretty", etc. These results highlight the importance of considering the original intent of the prompting and suggest that cultural practices on platforms such as Discord should be considered when designing interfaces that promote exploration and fair representation.

Bibliography

Jääskeläinen, P., & Åsberg, C. (2024, May). What’s the Look of" Negative Gender" and “Max Ethnicity” in AI-Generated Images? A Critical Visual Analysis of the Intersectional Politics of Portrayal. In Extended Abstracts of the CHI Conference on Human Factors in Computing Systems (pp. 1-9).

Sham, A. H., Aktas, K., Rizhinashvili, D., Kuklianov, D., Alisinanoglu, F., Ofodile, I., ... & Anbarjafari, G. (2023). Ethical AI in facial expression analysis: racial bias. Signal, Image and Video Processing, 17(2), 399-406.

Torricelli, M., Martino, M., Baronchelli, A., & Aiello, L. M. (2024, May). The role of interface design on prompt-mediated creativity in Generative AI. In Proceedings of the 16th ACM Web Science Conference (pp. 235-240).

Federico Aurora, Damir Nedic, Asgeir Nesøen, Dag Trygve Truslew Haug
University of Oslo, Norway

Exporting Mycenaean: from a Relational Database to EpiDoc XML files (and back again?)

Thursday, 06/Mar/2025 2:20pm - 2:40pm
ID: 276 / Session SP 06: 2
Short paper (full-text) | 15-minute presentation with a 5-minute Q&A
Keywords: Mycenaean Greek, Relational Databases, EpiDoc, TEI, Digital Editions

DAMOS is a (MariaDB) relational database, which contains all the published Mycenaean documents. These are administrative documents which constitute the earliest attestation (ca. 1400-1150 BCE) of the Ancient Greek language. The data in DAMOS can be accessed, searched and browsed online since the launch of its online interface in 2013. In this paper, after a general introduction on the data and the database structure, we describe the export function which we recently added to DAMOS and its online interface. This allows users to export data from the database as Epidoc-TEI compliant XML files. EpiDoc is an international, standard subset of TEI, used for the digital edition of documentary Ancient Greek and Latin texts, currently expanding to other epigraphic traditions (e.g., runes, ogham). The present work builds also on the collaboration with the EpiDoc community on the adaptation of EpiDoc to suitably represent the Mycenaean documents.

The aim of the EpiDoc export function is, thus, in compliance with the FAIR data principles, to increase the integration of the Mycenaean data with other similar datasets, to provide a sustainable format for the long time archiving of the DAMOS data, and to make it easier to reuse these data.

Finally, we briefly discuss the planned development of an import function, which would enable users to contribute data to DAMOS by uploading EpiDoc-XML files with new or revised versions of the Mycenaean texts, thus allowing, and fostering, collaborative work on the digital edition of the Mycenaean documents.

Narges Azizifard, **Lidia Pivovarova, Aatu Liimatta, Emily Öhman, Eetu Mäkelä**
University of Helsinki, Finland

Exploring Book Genres and Users’ Reviews on Reddit: A Pre-, During, and Post-COVID-19 Lockdowns Analysis

Wednesday, 05/Mar/2025 5:20pm - 5:40pm
ID: 208 / Session SP 03: 5
Short paper (abstract) | 15-minute presentation with a 5-minute Q&A
Keywords: Books, COVID-19, Reddit, Online Platforms, Text Analysis

Numerous online platforms, such as Goodreads, LovelyBooks, and Wattpad, offer features that enable users to create book reviews, comment on texts, engage in book discussions, and share recommendations (Rebora et al., 2021). One of the primary advantages of these platforms is their ability to provide succinct yet detailed summaries of well-known literary works, making them both accessible and informative for a wide audience (Agrawal, 2023). Furthermore, user reviews often highlight distinct writing styles, which become key points of discussion (Walsh & Antoniak, 2021).

Employing computational methods, such as text, sentiment, and emotion analysis (Jänicke et al., 2015; Pfahler et al., 2018), to analyze such online data can effectively capture the core concepts of books, as well as the main genres, strengths, and weaknesses identified by different users. This approach can help individuals make informed decisions about which books align best with their preferences. Additionally, computational analysis of large textual datasets can yield insights unattainable through traditional qualitative or hermeneutical methods (Schmidt, Kaindl, & Wolff, 2020).

In this study, we examine Reddit comments from two prominent subreddits, "books" and "suggestmeabook," across three time periods: before (April–September 2019), during (April–September 2020 and 2021), and after (April–September 2022) the COVID-19 lockdowns. The COVID-19 pandemic led to increased levels of various mental health challenges, largely due to sudden changes in daily life, such as social isolation caused by lockdowns (Zhu et al., 2021). Therefore, it is crucial to understand the evolving themes and discussions surrounding the impacts of COVID-19. Our research aims to investigate the extent to which the COVID-19 lockdowns influenced Reddit users' perceptions of books. Specifically, we seek to determine whether users expressed negative outcomes related to the lockdowns in their comments (posts) and how they defined, discussed, and debated books during the three distinct periods: before, during, and after the lockdowns.

By employing computational methods such as topic modeling, text, sentiment, and Plutchik emotion analysis (Plutchik, 1980), we aim to uncover various dimensions of book-related discussions. These methods will allow us to analyze the content and structure of user comments, revealing insights into how the pandemic has shaped literary preferences and conversations. Additionally, we seek to identify patterns in how users engage with literary content across these distinct phases. For example, we will analyze various characteristics of the books mentioned in submissions and their corresponding comments, including genre, date of publication, and author. We will also identify the co-occurrence of books across submissions and subsequent comments, and extract emotions and sentiments from the comments to explore any differences across the three time periods. By taking this approach, we aim to uncover a more profound and nuanced understanding of the relationship between social circumstances and literary engagement, particularly in how external factors shape the ways individuals interact with literature. Through this comprehensive exploration, we hope to shed light on various dimensions of this interaction, including how specific social contexts, such as the recent global pandemic, influence reading habits, behaviors, and preferences. Ultimately, this research will offer valuable insights into the broader impact of the pandemic on reading trends, particularly within the unique and diverse Reddit community.

Bibliography

Rebora, S., Boot, P., Pianzola, F., Gasser, B., Herrmann, JB., Kraxenberger, M., Kuijpers, MM., Lauer, G., Lendvai, P., Messerli, T.C. & Sorrentino, P. (2021). Digital humanities and digital social reading, Digital Scholarship in the Humanities, Volume 36, Issue Supplement_2, Pages ii230–ii250.

Agrawal, S. (2023). Twitterature: A New Digital Literary Genre. Evolutionary Studies in Imaginative Culture:73-80.

Walsh, M. & Antoniak, M., (2021). The Goodreads ‘Classics’: A Computational Study of Readers, Amazon, and Crowdsourced Amateur Criticism. Journal of Cultural Analytics 6 (2).

Jänicke, S., Franzini, G., Cheema, M.F. & Scheuermann, G. (2015). On close and distant reading in digital humanities: A survey and future challenges. In: EuroVis (STARs). pp. 83–103.

Pfahler, L., Elwert, F., Tabti, S., Morik, K. & Krech, V. (2018). What do you do with 5 million posts? versuche zum distant reading religioser online-foren. In: Vogeler, G. (ed.) Book of Abstracts, DHd 2018. pp. 335–338. Cologne, Germany.

Schmidt, T., Kaindl, F. & Wolff, C. (2020). Distant Reading of Religious Online Communities: A Case Study for Three Religious Forums on Reddit. In Proceedings of the Digital Humanities in the Nordic Countries 5th Conference (DHN 2020) (pp. 157-172). Riga, Latvia.

Zhu, J., Yalamanchi, N., Jin, R., Kenne, DR. & Phan, N. (2023). Investigating COVID-19’s Impact on Mental Health: Trend and Thematic Analysis of Reddit Users’ Discourse. J Med Internet Res ;25:e46867.

Plutchik, R. (1980). A general psychoevolutionary theory of emotion. Theories of emotion, 1:3–31.

Anda Baklāne, **Valdis Saulespurēns**
National Library of Latvia, Latvia

Extracting semantically related concepts from the corpus of Latvian novels: A comparative analysis of Word2Vec, GPT-4o, and Gemini-1,5 results

Wednesday, 05/Mar/2025 4:40pm - 5:00pm
ID: 218 / Session SP 02: 3
Short paper (abstract) | 15-minute presentation with a 5-minute Q&A
Keywords: Literary corpora, concept mining, Word2Vec, LLMs, transport in literature

In the field of concept mining within literary and historical corpora, a common task is identifying semantically related concepts (terms with similar meanings) or terms that fall under the same category or topic. This approach has a wide range of applications, from investigating specific concepts to analyzing broader semantic shifts in language and tracking the changing popularity of topics over time. Traditionally, manual or semi-automated methods have been used to discover or annotate concepts of interest. However, in recent years, researchers have increasingly scaled up their efforts by first adopting word embedding-based techniques and, more recently, zero-shot large language models (LLMs).

The goal of this paper is to compare the results of the Word2Vec and zero-shot LLMs results in identifying the terms related to transportation in the corpus of early Latvian novels. For the evaluation, the GPT-4o version and Gemini-1,5 were used.

It is recognized that large language models perform well in the tasks of categorizing and annotation that were previously realized manually or by using supervised machine learning means (including identifying named entities, semantically related, or domain-specific concepts) (Karjus, 2024; Fan et al., 2023; Ziems et al., 2023). The validation and quality control of the results generated LLMs, however, present significant challenges; additionally, there are concerns regarding the accessibility of these models. Currently, the performance of language models is notably less effective for smaller languages, such as Latvian, and the usability of the methodologies employed requires further validation at this stage of model development.

During preliminary research, we identified several current state-of-the-art models that provided acceptable performance for the Latvian language. Out of those models, we selected the better-known and better-supported GPT-4o and Gemini-1.5 based models. Although non-commercial models would be preferable for research purposes, currently available fully open models do not provide satisfactory results.

The authors of this paper have previously explored the use of Word2Vec word embeddings to automatically identify and extract (and subsequently quantitize) terms related to specific categories or semantic domains. In a case study, we examined concepts related to urban transportation within a corpus of Latvian early novels (1879–1940) (LatSenRom) (Kristsone et al., 2024). The Word2Vec model was trained on 457 novels, along with an additional 172,240 articles from historical periodicals from 1920 to 1940. The resulting embeddings enabled the identification of terms related to urban transport vehicles, as well as variants of different spellings of words, which arose either from changes in orthography over time or from errors in optical character recognition. This allowed us to considerably increase the number of found terms compared to working with subjectively devised lists and lexicons of vehicles.

In extracting a comprehensive list of vehicle terms using Word2Vec embeddings, an initial subjective list of vehicles was compiled, and queries based on these keywords were used to generate broader lists of semantically similar concepts. This approach identified several dozen types of land vehicles, including various horse-drawn carriages, mechanized vehicles, and specific car brands. In the current paper, the authors expand the range of transportation domain by adding air and water vehicles.

To explore the results of the GPT-4o and Gemini-1,5 models, a workflow was devised to query individual novels to retrieve terms related to transportation; subsequently, the results were summarized to compare with the totality of results for LatSenRom retrieved from the Word2Vec model. The workflow was run in batch mode, using APIs provided by respective model holders – OpenAi and Google.

Several types of prompts were used for the exploration, including queries in Latvian and English, prompts based on keywords mimicking the workflow used for Word2Vec, and other types of questions. As LLMs are prone to various anomalies and hallucinations, a hybrid approach was utilized to verify results and minimize false positives. Different instructions were used to retrieve normalized terms or various forms treated as "anomalies" by the models; these terms are relevant for further reuse of the lists of words (e.g., quantification based on the mentions over time).

Bibliography

Fan, Yaxin, et al. 2023. "Uncovering the potential of chatgpt for discourse analysis in dialogue: An empirical study." In arXiv:2305.08391. Pre-published.

Karjus, Andres. 2023. "Machine-assisted mixed methods: augmenting humanities and social sciences with artificial intelligence." Computer Science. Computation and Language. arXiv:2309.14379 [cs.CL]. https://doi.org/10.48550/arXiv.2309.14379. Pre-published.

Kristsone, Eva-Eglāja, Anda Baklāne, and Valdis Saulespurēns. 2024. "Pilsētas transporta līdzekļi latviešu senākajos romānos." Latvijas Nacionālās bibliotēkas Zinātniskie raksti, Nr.12 (XXXII): Tuvlasījums/ Tāllasījums. In press.

Mikolov, Tomas, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. "Efficient estimation of word representations in vector space." In arXiv:1301.3781v3 [cs.CL]. https://doi.org/10.48550/arXiv.1301.3781

Törnberg, Petter (2023). "ChatGPT-4 outperforms experts and crowd workers in annotating Political Twitter Messages with Zero-Shot Learning." In 10.48550/arXiv.2304.06588. In http://arxiv.org/abs/2304.06588. Pre-published.

Törnberg, Petter (2024). "Best practices for text annotation with large language models." In http://arxiv.org/abs/ 2402.05129. Pre-published.

Ziems, Calet, William Held, Omar Shaikh, Jiaao Chen, Zhehao Zhang, and Diyi Yang. 2023. "Can Large Language Models Transform Computational Social Science?" In Computational Linguistics, pp.1-53. 10.1162/coli_a_00502

Katrine F. Baunvig
Aarhus University, Denmark

From Numbers to Narratives - Word Embeddings and Semantic Graphs as Hermeneutic Strategies in Textual Analysis

ID: 310 / WS04A: 3
Explorations of the dynamics of cultural phenomena in text corpora
Keywords: Word Embedding; Semantic Graphs; Quantitative Textual Analysis; Qualitative Textual Analysis

The conventional divide between qualitative and quantitative textual analysis—where the former delves into interpretive nuance and the latter focuses on scalable computation—is far less definitive than it may seem. In this workshop tutorial, we explore how neural word embeddings and semantic graphs, often emblematic of quantitative methodologies, can be reconceptualized as hermeneutic tools that bridge this dichotomy. Using the monumental corpus of N.F.S. Grundtvig (1783-1872), a Danish polymath whose works profoundly shaped national identity and religious discourse, we demonstrate how computational approaches can complement and deepen interpretive scholarship (Baunvig & Nielbo 2023).

We present two interwoven arguments. First, that the qualitative resides within the quantitative (Baunvig 2024). Computational methods like word embeddings not only quantify textual relationships but also require interpretive decisions at every stage—from corpus selection and preprocessing to the visualization of semantic spaces. These choices, far from neutral, imbue the analysis with interpretive depth, revealing that computation itself is a deeply hermeneutic act.

Second, meaningful computational analysis depends on an intimate understanding of the dataset and its cultural and historical contexts (cf. Nielbo, Baunvig & Gao 2018). Why analyze an entire authorship like Grundtvig's? What conditions render such a project significant? His corpus, spanning more than a thousand works, offers a unique vantage point into the intersections of Danish cultural memory and religious imagination. Yet its richness also necessitates careful reflection on the alignment of research questions with the dataset's scope and the ethical considerations underpinning such scholarly endeavors.

This workshop offers participants practical insights into leveraging word embeddings and semantic graphs to uncover symbolic structures across texts. It underscores the necessity of balancing technical rigor with interpretive sensitivity, cultivating an approach that is as methodologically sound as it is intellectually meaningful. By reframing quantitative tools as extensions of hermeneutic inquiry, we invite participants to transcend traditional boundaries and reimagine the possibilities of computational textual analysis.

Join us as we navigate this compelling intersection of numbers and narratives, illuminating how data-driven methods can not only augment but also transform the interpretive study of texts.

Bibliography

Baunvig, KF, 2024, Grundtvig og monstrene, Odense: University Press of Southern Denmark.

Baunvig, KF & Nielbo, KL 2023, "Benign Structures: The Worldview of Danish National Poet, Pastor, and Politician N.F.S. Grundtvig (1783-1872)". In A Rockenberger, J Tiemann & S Gilbert (red), Conference Proceedings : DHNB2023, 1, Oslo University Press, Oslo, Digital Humanities in the Nordic and Baltic Countries Publications, No. 1, vol. 5, p. 1-10, 7th Conference Digital Humanities in the Nordic and Baltic Countries, Oslo, Norge, 08/03/2023. https://doi.org/10.5617/dhnbpub.10646

Baunvig, KF 2023, A Computational Future? Distant Reading in the Historical Study of Religion. i M Freudenberg, F Elwert, T Karis, M Radermacher & J Schlamelcher (red), Stepping Back and Looking Ahead: Twelve Years of Studying Religious Contact at the Käte Hamburger Kolleg Bochum. Brill, Leiden, Dynamics in the History of Religions, vol. 13, p. 325-352. https://doi.org/10.1163/9789004549319_013

Nielbo, KL, Liu, B, Baunvig, KF & Gao, J 2018, "A curious case of entropic decay: Persistent complexity in textual cultural heritage", Digital Scholarship in the Humanities, bind 34, nr. 3, p. 542-557. https://doi.org/10.1093/llc/fqy054

Katrine F. Baunvig, Jon Tafdrup, Kirsten Vad, Krista Stinne Greve Rasmussen
Aarhus University, Denmark

From Bias to Insight: Computational Challenges and Opportunities in the Humanities

Thursday, 06/Mar/2025 9:00am - 9:30am
ID: 198 / Session LP 05: 1
Long paper (abstract) | 20-minute presentation with a 10-minute Q&A
Keywords: Bias in Historical Data; Digital Scholarly Editions (DSE); Computational Humanities; Data Preservation; Cultural Canonization

1. Introduction: Bias as a Critical Pathway to Historical Insight

The integration of computational methods into traditional humanities disciplines has transitioned from a field of potential to one central to the transformation of concrete research practices. These advancements have significantly reshaped how scholars interpret cultural data. However, with these advancements come challenges. In this paper, we address the challenge of transforming bias—often regarded as a flaw—into a valuable analytical tool. In historical datasets, bias is not simply a problem to be eliminated but a reflection of the cultural, political, and social forces that shaped the material. This understanding shifts bias from an obstacle to a critical pathway for uncovering deeper meanings and insights within given data.

In other words, this paper explores how computational methods can harness the biases present in historical corpora to produce more nuanced interpretations. Moreover, it addresses the role of data preservation strategies in ensuring that these biases are not lost to future scholarship. Using the Digital Scholarly Edition (DSE) of *Grundtvig's Works* as a case study, we seek to demonstrate how bias can be both a challenge and an interpretative opportunity in the development and sustainability of digital archives. Within the framework of the DSE, the emphasis in this case is on a single authorship, reflecting both the inherent bias of one individual from a particular era and the bias inherent in the DSE itself.

2. Preserving Bias for Future Insight: Data Quality and Longevity

The preservation of historical data is not merely a technical concern but a philosophical one that directly impacts how future scholars will interpret cultural data. Traditional humanities research, which often centers on close readings of carefully curated texts, allows for a high degree of control over the sources. In contrast, large-scale computational projects must contend with the digitization of incomplete, inconsistent, or biased texts, introducing challenges in maintaining the quality of the data while preserving its interpretative richness.

The *Grundtvig's Works* DSE exemplifies the effort needed to produce high-quality digital editions that acknowledge, rather than erase, the biases inherent in historical data (cf., Oltmanns et al., 2019; Pierazzo, 2014). While rigorous data cleaning and curation are essential, questions of long-term preservation loom large. Without sustainable infrastructures and open data standards, digital scholarly editions risk becoming obsolete or inaccessible, leading to what some scholars have called a potential 'digital wasteland' (Baunvig et al., 2023). Preservation is crucial not just for maintaining access to the content, but for safeguarding the biases within the data that reflect the ideological and socio-political conditions of the time.

Rather than prioritizing rendition—where texts are cleaned and 'corrected' for readability—digital humanists must emphasize storage and preservation to ensure that future scholars can continue to engage with the historical biases embedded within the texts. This allows for a more reflective approach to the past, where bias is seen not as an error to be erased, but as a feature to be explored.

3. Bias as an Interpretative Lens in Historical Datasets

Bias is an inherent concern in any humanities project, but it takes on greater complexity in large-scale computational work. Traditional humanistic methods give scholars a degree of control and intentionality over text selection, but computational projects often involve datasets digitized at scale, where biases are more deeply embedded and sometimes amplified. The digitization process itself can introduce errors, but more critically, it can reflect and perpetuate historical biases from the time when the texts were created. These biases may concern the ideological, religious, or nationalist frameworks of the authors and societies in question. For example, in the *Grundtvig's Works* DSE, the digitized materials reflect nationalist and religious assumptions of 19th-century Denmark, giving scholars insight into the cultural forces shaping those works (Rasmussen et al., 2022). Bias, in this context, becomes a rich source of information, providing pathways for the dominant narratives of the past – and for exploring exclusion, and marginalization.

Computational approaches provide opportunities for identifying and analyzing these biases at scale, while humanistic reflection remains essential. Digital humanists must balance computational rigor with interpretive depth, recognizing that bias detection algorithms, often used to ensure fairness in contemporary datasets, can be repurposed to trace historical ideologies in ways that enhance our understanding of the past.

4. Refining Bias Detection Algorithms for Historical Data

Bias detection algorithms, initially designed to ensure fairness in modern datasets, can be refined to address the unique challenges posed by historical corpora. By refining bias detection algorithms, digital humanists can preserve these interpretive complexities while analyzing historical data. One refinement involves making the algorithms *context sensitive*. Rather than applying contemporary standards of fairness, algorithms tailored for historical datasets would flag specific biases related to gender, race, or religion as elements to explore, not to remove. This allows scholars to critically engage with the biases in their historical context, providing a more nuanced understanding of the data.

Additionally, biases in historical texts are often *multilayered*, involving not only individual authors but also broader institutional and societal forces. Algorithms refined to detect these multiple layers would help scholars uncover both explicit biases and the silences or omissions that reveal underlying patterns of exclusion. Biases also evolve over time, reflecting changing cultural norms. Algorithms capable of tracking these temporal dynamics would allow scholars to explore how specific biases shifted across different historical periods, offering new insights into the cultural transformations of the time. Likewise, cultural specificity is key—what constitutes bias in one historical or regional context may not apply in another. By training algorithms on region-specific datasets, scholars can enhance their ability to detect biases unique to the cultures they are studying.

Another refinement lies in the use of *bias as a tool for reflection*. Algorithms can be designed to not only detect bias but also help scholars interpret how these biases relate to the broader socio-political structures of the time. This allows for a deeper exploration of power dynamics embedded within the texts. User-directed features, where researchers can specify the types of biases they wish to explore, would further enhance the utility of these tools for historical analysis. Finally, advanced visualization tools can be integrated into bias detection algorithms, offering visual maps of where and how biases appear within a corpus. These visualizations allow for a more interpretive engagement with the data, highlighting intersections of different biases and their shifts over time, making the complexity of historical corpora more accessible to scholars.

5. Ethical Considerations: A Reflective Approach to Bias

As computational methods become more prevalent, ethical considerations regarding the treatment of bias in historical datasets become critical. The biases in cultural, political, and ideological contexts are not simply flaws to be corrected but essential features for understanding the conditions under which historical texts were created.

Rather than erasing these biases in the pursuit of neutrality, scholars should aim to understand how they shaped the historical record. By embedding ethical reflection into every stage of the research process, digital humanists can ensure that computational methods do not merely amplify dominant narratives but also reveal marginalized voices, offering new insights into the power dynamics of the past. Ethical frameworks specific to the digital humanities are needed to engage more deeply with these complexities and to ensure that computational tools are used in a way that enriches, rather than diminishes, the interpretative potential of historical data (Kleinberg et al., 2016).

6. Canonical Bias in *the Grundtvig's Works* DSE: A Paradox of Cultural Preservation

Finally, in the context of bias as both challenge and opportunity, the DSE of *Grundtvig's Works* presents a compelling paradox. On one hand, it provides scholars with high-quality, meticulously annotated texts that enhance our understanding of Grundtvig's substantial impact on Danish national identity, religious thought, and cultural history. On the other hand, the very creation of the DSE is shaped by a significant form of bias—one embedded in the processes of cultural canonization and national self-conception.

Grundtvig's prominence in Danish history has ensured that he is not just an important figure, but a cultural icon, occasionally referred to as a 'cultural saint', for his role in shaping the nation's democracy, educational system, and church life. It is this canonization that has enabled the DSE project to attract significant funding, with (currently) well over 150 million DKK invested in making Grundtvig's extensive writings available to the public and scholars alike. This level of financial and institutional support is not distributed equally across all figures or texts from Danish history. Instead, it is directly tied to Grundtvig's symbolic status as a central figure in the collective Danish imagination.

Here, we encounter a form of *institutional bias* that, while making Grundtvig's works more accessible, simultaneously perpetuates the focus on canonical figures. The DSE was made possible not simply because of the scholarly value of Grundtvig's writings, but because of the cultural weight his name carries. This raises critical questions about whose works are considered worthy of preservation and extensive study, and which voices remain marginalized or overlooked. In this sense, the bias embedded in the DSE reflects broader patterns of historical selection, where resources are allocated based on cultural prominence rather than equitable representation.

However, this paradoxical bias also opens up an opportunity for reflection. The DSE's existence, made possible by Grundtvig's cultural stature, allows us to interrogate the very processes of canonization and the role of bias in shaping scholarly endeavors. The creation of such a comprehensive, well-funded project demonstrates how bias operates not only within the content of historical texts but also within the systems of support that determine which texts are preserved, studied, and celebrated. Moreover, this bias is not static. Just as computational tools can reveal and analyze the biases within historical data, so too can the systems of cultural preservation evolve. The attention paid to canonical figures like Grundtvig is beginning to shift, with increasing efforts to broaden the scope of cultural heritage projects to include a more diverse range of voices and materials. While the DSE represents a pinnacle of canon-focused scholarship, it also exemplifies how bias—paradoxically—serves both to sustain and to challenge the boundaries of cultural memory. By critically engaging with this bias, we can better understand the socio-political forces that shape the preservation of historical data. In the case of the DSE, this awareness prompts us to reflect on how the mechanisms that prioritize certain figures might, in the future, be reconfigured to offer a more inclusive, socially sustainable approach to cultural heritage.

7. Conclusion: Embracing Bias as Insight in Digital Scholarship

The integration of computational methods into the humanities marks the beginning of a new era—one where technology amplifies, rather than replaces, traditional humanistic inquiry. The *Grundtvig's Works* DSE serves as a powerful example of how computational methods can open new interpretive avenues by addressing the challenges of bias, data quality, and preservation.

Computation should not be viewed as a tool for efficiency but as a key to deeper inquiry, particularly in its ability to expose biases in historical datasets that might otherwise remain hidden. Bias, in this context, becomes not a flaw to be eliminated but a critical aspect of historical interpretation. Digital humanists must leverage these biases to illuminate the socio-political forces that shaped historical texts, gaining deeper insights into the past.

The future of digital scholarship lies in maintaining a balance between the interpretive complexity of humanistic traditions and the power of computational tools. By fostering interdisciplinary collaboration, developing robust ethical frameworks, and building sustainable infrastructures, the humanities are poised to thrive in the digital age, producing richer and more nuanced understandings of our cultural heritage.

Bibliography

Baunvig, K.F., Rasmussen, K.S.G., Møldrup-Dalum, P., & Vad, K., 2023, Storage Over Rendition. Towards a Sustainable Infrastructure in the Digital Textual Heritage Sector. Digital Humanities in the Nordic and Baltic Countries Publications 5(1): 24047. https://doi.org/10.5617/dhnbpub.10667.

Friedman, B., & Nissenbaum, H., 1996, "Bias in computer systems", ACM Transactions on Infor-mation Systems (TOIS), 14(3), 330-347.

Kleinberg, J., Mullainathan, S., & Raghavan, M., 2016, "Inherent trade-offs in the fair determination of risk scores", arXiv preprint arXiv:1609.05807.

Oltmanns, E., Hasler, T., Peters-Kottig, W., & Kuper, H.-G., 2019, Different Preservation Levels: The Case of Scholarly Digital Editions. Data Science Journal, 18(1), 51.
Pierazzo, E., 2015, Digital Scholarly Editing: Theories, Models and Methods, Ashgate.

Rasmussen, K.S.G., Tafdrup, J., Ravn, K.S., & Baunvig, K.F., 2022, The Case for Scholarly Edi-tions. Proceedings of the 6th Digital Humanities in the Nordic and Baltic Countries Conference (DHNB 2022), pp. 401-405. CEUR Workshop Proceedings, Vol. 3232. https://ceur-ws.org/Vol-3232/paper39.pdf.

Joanna Beaufoy[1], **Lars Kjær**[2]
[1]University of Copenhagen, Denmark; [2]The Royal Danish Library

Assembling artificial light and literary emotions: computational approaches to literary lighting in nineteenth-century literature set in Paris

Wednesday, 05/Mar/2025 5:20pm - 5:40pm
ID: 145 / Session SP 04: 5
Long paper (abstract) | 20-minute presentation with a 10-minute Q&A
Keywords: automatic close reading, ChatGPT, artificial light, history of the emotions, nineteenth-century French literature

A candle flickers, burning low, as two lovers rush into bed (Zola 1890, 244).[1] A troubled figure paces up and down a pavement, looking up at a shadow at a lit window (Flaubert 1869, 39).[2] As a train slows down entering Paris, the flood of electric light dazzles a passenger and stops him mid-sentence (Huysmans 1896, 459).[3] In literature — as in theatre, cinema and urban design — artificial light is a narrative and stylistic tool involved in representing emotional states. Literary lighting, or literary illumination as it has been called by Leahy (2019),[4] is most immediately recognisable in the form of literary metaphors such as the *coup de foudre*: the lightning bolt of falling in love, or the art historical motif of a lover appearing as if in a glow of light. Artificial lighting is not however limited to metaphor; rather, it is a technique of realism. Guiding the reader's 'sight' in a construction of verisimilitude, it functions as part of what Barthes called the 'reality effect',[5]creating an impression of reality with this literary style perfected in the nineteenth century.

The mid- to late-nineteenth century was the era of the most radical developments in artificial lighting technology in history. Paris was for this brief period the main stage of experimentation in urban lighting, and visitors would travel from around the world to see the city lighting. This was also a period of fervent literary production, and the development of realist and naturalist techniques that engaged with the material world in unprecedented ways. We are interested in literary lighting historically, asking in what ways the transformations in lighting technologies during the nineteenth century, spanning the eras of oil, gas and electric light, found their way into literature. Our (human) reading observes the frequency of artificial light's collocation with depictions of romantic emotions, which led us to want to explore this on a greater scale.

The three examples from literature above were all found using artificial intelligence. In our paper, we demonstrate how automated reading can successfully find examples where writers were using lighting techniques to represent emotional states, focusing on those adjacent to love and romance. The history of the emotions is the object of many academic dreams: how can we ever know what was felt by our predecessors? The question of how the new lighting technologies of gas light and electric light in the mid-late nineteenth century correlated with literary evocations of emotional states is complex, but we demonstrate that computational methods can go some way onto shedding light on the problem. Our paper presents a handmade corpus of ninety novels by canonical authors such as Balzac, Flaubert, Colette, Zola, Rachilde, Maupassant, and Proust, as well as lesser-known writers, publishing novels set in Paris between 1841 and 1928. Ninety novels, 9 850 523 words, is hardly 'big data' in data processing terms, but it is so in the context of the conventions of literary studies.

French literary studies is an established field with a long research tradition, and the work of canonical writers has not previously been analysed in any published work using the Large Language Models afforded by artificial intelligence. Because of their wealth of material describing and evoking emotional states, works of literature are valuable sources when we seek to collocate the history of technology with the history of the emotions. The experiments in our paper are presented within a broader critique of artificial intelligence's usefulness to the field of literary scholarship. In the paper, we discuss their concrete implications for research practices, and suggest opportunities for more tailor-made resources for literary scholars based on this case study.

Our approach has been to experiment with various digital methods before arriving at the method we could use to shed light on our research question, and by presenting these early experiments in our paper we make our path of reasoning visible. We briefly outline our initial experiments and highlights our difficulties in identifying patterns using known NLP methods, such as word lists, word frequencies, concordance, and n-grams.

Since OpenAI introduced ChatGPT, we believe that new opportunities have emerged for utilizing technology within the humanities, and we thought it would be natural to incorporate ChatGPT and explore its potential in relation to the known NLP methods.

The method that has given us the best results , therefore, is based on the Python libraries LangChain (Chase, H. (2022*)*[6] and the OpenAI API (OpenAI 2024).[7] The inspiration to use these libraries comes from short courses at Deeplearning.AI, an education company specialized in technology (Deeplearning.ai 2024).[8] These results are the main presentation of our paper.

We use the functions of LangChain and the OpenAI API for preparation and search. The preparation involves splitting our corpus into text chunks and embedding and vectorizing these chunks. The embedding technology allows for semantic searches (Kublik, S., & Saboo, S. 2023, 37)[9]. This way, we can search, retrieve, and evaluate whether the different text chunks contain examples that we can use in our study. The evaluation of text chunks is based on GPT's ability to perform zero-shot classification, which means that the model can classify without fine-tuning (Rothman, D. 2022, 170).[10]

The results we obtain from applying LangChain, the OpenAI API, and OpenAI's GPT-4 omni model are of a different nature compared to the results we have previously achieved using known NLP methods. In the past, we have spent a lot of time analyzing the distribution of selected keywords and n-grams to observe patterns. As we show in this study, one of the contributions of LLMs to digital humanities research can be examining the contexts around keywords.

In contrast to earlier NLP methods, the LLM can identify if the keyword is part of a context that constitutes a semantic field of interest. Specifically, this is relevant for asking questions related to emotional states, which cannot be resumed in a list of words. This is why, after Santos (2023), we refer to our reading practice in this paper as automatic close reading, rather than distant reading.[11]

Using our approach, we demonstrate in the paper how computational reading and human reading can be used together, the former to find the examples in seconds, the latter to explore them with the knowledge of the literary era in particular, and the literary knowledge in general, required to understand the relationships between the characters and the effects of artificial light upon them. Further research could hone the asking of questions to AI assistants in a literary context, to find the ceiling of complexity or subtlety the technology is currently capable of. These capabilities will increase with time. Considering that only

1.01965% of the documents for the development of gpt-3 were French, there is still a lot of potential there (Brown, Tom B. et al. 2020)[12].

The paper's main contribution to the field is therefore to provide a practical case study where recent computational tools are applied to a nineteenth-century French literature corpus, and to demonstrate the strengths and weaknesses of these tools for literary scholarship. By applying solid computational methods to the question of artificial lighting and emotional states, we can suggest that artificial lighting developments may have had an impact on realist techniques, and we illustrate this with examples from the work of Flaubert and Zola. Our results show how the arrival of gas lighting and then electricity in Paris during the mid nineteenth century to the early twentieth century had its literary reflection in how writers used artificial lighting to depict loving or romantic emotions. Our paper suggests that computational methods to conduct automated close reading can provide significant insights into changes over time: by quantifying these examples, we are able to suggest using a graph how writers' engagement with different lighting technologies developed over time between 1841 and 1928.

The paper makes available a method and Python code we have devised that finds and displays examples of nineteenth-century novelists' use of artificial light to articulate loving feelings in works set in Paris. More broadly, the results produced by the machine, and our human analyses of these, are relevant to scholars interested in the interactions between technology and emotion, both in a literary and in the urban worlds we now inhabit.

[1] Zola, Emile. 1890. *La Bête humaine*, effective 14 October, 2024, https://fr.wikisource.org/wiki/La_Bête_humaine/Texte_entier

[2] Flaubert, Gustave. 1869. *L'Education sentimentale* effective 14 October, 2024, https://fr.wikisource.org/wiki/L'Éducation_sentimentale,_éd._Charpentier,_1891/Texte_entier

[3] Huysmans, Joris-Karl. 1896. *En route*, effective 14 October 2024, https://fr.wikisource.org/wiki/En_route_(Huysmans)/Texte_entier

[4] Leahy, Richard. 2018. *Literary Illumination: The Evolution of Artificial Light in Nineteenth-Century Literature*. Cardiff: University of Wales Press.

[5] Barthes, Roland. 1968. 'L'effet du réel', *Communications*, n° 11, 1968 passage=84-89 (DOI 10.3406/comm.1968.1158).

[6] Chase, H. 2022. LangChain [Computer software]. Effective October 13, 2024, https://github.com/langchain-ai/langchain)

[7] OpenAI 2024, OpenAI Python API library, Effective October 13, 2024, https://pypi.org/project/openai/

[8] Deeplearning.ai. 2024. Effective October 13, 2024, https://www.deeplearning.ai/

[9] Kublik, S., & Saboo, S. 2023. Gpt-3 : The ultimate guide to building nlp products with openai api. Packt Publishing, Limited.

[10] Rothman, D. 2022. Transformers for natural language processing: Build, train, and fine-tune deep neural network architectures for nlp with python, hugging face, and openai's gpt-3, chatgpt, and gpt-4. Packt Publishing, Limited.

[11] Santos, Diana. 2023. n.d. 'Literature Studies in Literateca: Between Digital Humanities and Corpus Linguistics'.

[12] Brown, Tom B. et al. 2020. "Language Models are Few-Shot Learners." arXiv, https://arxiv.org/abs/2005.14165. Effective October 13, 2024, gpt-3/dataset_statistics/languages_by_document_count.csv at master · openai/gpt-3 · GitHub

Jonathan Westin, Dorna Behdadi, Daniel Brodén, Mats Fridlund
Gothenburg Research Infrastructure in Digital Humanities, Dept. of Literature, History of Ideas and Religion at the University of Gothenburg, Sweden

AI in the GLAM sector - Opportunities and Challenges

ID: 254 / **Panel 4: 1**
60-minute panel
Keywords: GLAM: cultural heritage: artificial intelligence: machine learning

Introduction

Aligned with the DHNB 2025 theme, 'Digital Dreams and Practices,' this panel aims to explore the transformative potential of Artificial Intelligence (AI) within the GLAM (Galleries, Libraries, Archives, and Museums) sector. As cultural heritage institutions continue to digitize collections and adopt new technologies, AI presents groundbreaking opportunities and challenges that warrant closer examination.

For over two decades, the GLAM sector has focused on digitizing and making collections accessible (Astle & Muir, 2002). Digitization involves converting analog materials into digital form, a process that requires critical intellectual and technical decisions, which affect the possibilities and constraints for knowledge dissemination and the sensory experiences that new digital artifacts enable (Dahlström, 2010; Westin, 2023). The introduction of AI into this process further complicates the role of technology as a mediator, especially when used for markup, analysis, or reconstruction of cultural heritage. The use of machine learning and AI to annotate digitized material has been criticized due to the high risk of perpetuating stereotypes, biases, and outdated explanatory models (Villaespesa & Murphy, 2021; Balbi & Calise, 2023; Huang & Liem, 2022; Dikow et al., 2023). This issue is exacerbated by the fact that physical archives and collections themselves may be rooted in various biases and outdated explanatory models in terms of their focus, methods of collection, sorting, and purpose.

AI technologies are already being implemented in the GLAM sector for tasks such as automating metadata creation, image recognition, and digital preservation. These developments promise to enhance the accessibility of collections and improve the efficiency of curatorial processes. For instance, AI-powered algorithms can aid in identifying previously overlooked patterns in historical documents or suggesting connections between seemingly unrelated items (Neudecker, 2022; Lee, 2023). However, these advancements also introduce concerns related to the loss of human expertise, the potential homogenization of cultural narratives, and ethical dilemmas in data handling. The intersection of AI with cultural heritage raises questions about the extent to which human curators should rely on automated systems for tasks traditionally governed by scholarly interpretation and emotional resonance (Boiano et al., 2024).

This panel will not only address the technical and operational implications of AI but also emphasize the need for a responsible approach to its integration. Ethical considerations, such as ensuring transparency in AI decision-making processes, safeguarding against algorithmic biases, and promoting inclusivity in the representation of diverse cultures and languages, are central to the conversation (Dignum, 2019). The panelists will collectively examine the role of AI in reshaping heritage practices, reflecting on how institutions can balance innovation with the preservation of cultural integrity.

Our panel brings together researchers and professionals from Digital Humanities, Data Science, Cultural Heritage, and the GLAM sector to share their experiences with AI in cultural heritage contexts. Through individual presentations and a joint discussion, we aim to critically assess the benefits and limitations of AI applications, while exploring how these technologies can be responsibly integrated into heritage practices.

Panel structure (90 minutes in total)

1. Introduction (10 minutes):

The moderator, Jonathan Westin, Associate Professor in Conservation and Deputy Director of Gothenburg Research Infrastructure in Digital Humanities (GRIDH) at the University of Gothenburg, will set the stage by providing an overview of the panel's theme and introducing key discussion points regarding the application of AI in the GLAM sector. His introduction will also touch on broader trends in digital heritage, situating the conversation within the ongoing digitization efforts of cultural institutions and the increasing reliance on AI technologies to tackle curatorial challenges, enhance accessibility, and transform public engagement

2. Individual presentations (5-7 minutes each):

Following the introduction, each panelist will deliver a short presentation reflecting on their practical experiences and insights into using AI within GLAM. These presentations will address the following questions (and others):

- What are the real-world applications of AI in your cultural heritage practice?
- What challenges have you encountered, such as ethical concerns or skills gaps?
- How can AI tools contribute to new forms of engagement with cultural heritage?
- How do you ensure accuracy and reliability of AI-generated annotations?
- What strategies do you employ to mitigate biases in AI applications?
- How do you balance the use of AI with traditional methods in your work?

3. Joint discussion:

The panelists will, together with the public, engage in a 45-minute discussion moderated by Jonathan Westin. The discussion will focus on issues such as the balance between automation and human expertise, the societal impact of AI in GLAM, and lessons learned from partnerships between academic researchers, memory institutions, and the public. Key themes will include how AI can both enhance and challenge existing curatorial practices, as well as strategies for fostering interdisciplinary collaboration in future AI initiatives.

Outlines of individual contributions

- **David Haskiya** (Swedish National Archives): Haskiya is Head of Unit for the AI lab and Data Services at the Swedish National Archives. Haskiya's presentation will discuss the intersection of AI and archival science, and focus on how AI tools are being implemented to enhance archival processes, improve accessibility, and derive new insights from historical collections.

- **Wilhelm Lagercrantz** (National Historical Museums): Lagercrantz is Operations Manager in Digital productions at the National Historical Museums. Lagercrantz' contribution will explore how national museums are adapting to and implementing digital solutions, focusing on the intersection between technological innovation and museum practice. Drawing from his experience at the National Historical Museums, he will discuss strategies for digital collection management, online accessibility, and public engagement through digital platforms.

- **Karin Glasemann** (ArkDes): Glasemann is the Head of Collections at ArkDes and the project manager for the research project "Ett nätverk av platser – Öppna länkade bebyggelsedata som forskningsinfrastruktur.

Bibliography

Astle, P. J., & Muir, A. (2002). Digitization and preservation in public libraries and archives. Journal of Librarianship and Information Science, 34(2), 67-79.

Balbi, C., & Calise, A. (2023). The (theoretical) elephant in the room. Overlooked assumptions in computer vision analysis of art images. Signata. Annales des sémiotiques/Annals of Semiotics, (14).

Dahlström, M. (2011). Critical editing and critical digitisation. In Text Comparison and Digital Creativity (pp. 77-97). Brill.

Dikow, R., DiPietro, C., Trizna, M., BredenbeckCorp, H., Bursell, M., Ekwealor, J., ... & White, A. (2023). Developing responsible AI practices at the Smithsonian Institution. ARPHA Preprints, 4, e113335.

Huang, H. Y., & Liem, C. C. (2022, June). Social inclusion in curated contexts: Insights from museum practices. In Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (pp. 300-309).

Villaespesa, E., & Murphy, O. (2021). This is not an apple! Benefits and challenges of applying computer vision to museum collections. Museum Management and Curatorship, 36(4), 362-383.

Westin, J. (2021). Arosenius Translated. Digitisation as a Rephrasing of Meaning. Nordisk Museologi, 31(1), 40-55.

Bibliography

Astle, P. J., & Muir, A. (2002). Digitization and preservation in public libraries and archives. Journal of Librarianship and Information Science, 34(2), 67-79.

Balbi, C., & Calise, A. (2023). The (theoretical) elephant in the room. Overlooked assumptions in computer vision analysis of art images. Signata. Annales des sémiotiques/Annals of Semiotics, (14).

Dahlström, M. (2011). Critical editing and critical digitisation. In Text Comparison and Digital Creativity (pp. 77-97). Brill.

Dikow, R., DiPietro, C., Trizna, M., BredenbeckCorp, H., Bursell, M., Ekwealor, J., ... & White, A. (2023). Developing responsible AI practices at the Smithsonian Institution. ARPHA Preprints, 4, e113335.

Huang, H. Y., & Liem, C. C. (2022, June). Social inclusion in curated contexts: Insights from museum practices. In Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (pp. 300-309).

Villaespesa, E., & Murphy, O. (2021). This is not an apple! Benefits and challenges of applying computer vision to museum collections. Museum Management and Curatorship, 36(4), 362-383.

Westin, J. (2021). Arosenius Translated. Digitisation as a Rephrasing of Meaning. Nordisk Museologi, 31(1), 40-55.

Viken Berberian
American University of Armenia (AUA), Armenia

Causality & Climate Change: a Comparison of AI-Generated and Student Written Essays

ID: 104 / Poster Session 2: 2
Poster and demo (abstract) with accompanying a 1-minute lightning talk
Keywords: Digital Humanities, Language, Rhetoric, Voyant Tools, AI in Higher Education, Cause and Effect Essay, Digital Scholarship, ChatGPT, Chatbot, Quantitative Writing Tools, AI in Writing, Qualitative Analysis, ChatGPT in Science, Pedagogical Practices

My poster and lightning talk will address the linguistic, rhetorical, stylistic, and research choices in essays written by 75 second-year expository writing students and compare them to those generated by a chatbot. Students will write a three-page essay on the causes and effects of climate change. The findings will reflect the work of English majors enrolled in an expository writing course at AUA, the American University of Armenia, comparing their work to those generated by ChatGPT. Criteria and methodology: Three sources will be cited and one of them will be peer reviewed. Each essay will consist of 5 paragraphs and focus on one cause, one effect, and conclude with one call to action. The essay will be limited to three pages. Students will have four weeks to research and write their essay alongside a ChatGBT-generated essay on climate change that will use the same criteria. The ChatGPT essay will be generated only after a student has researched and written her essay without the use of a chatbot. Students will submit two versions of their essay: one written by them; the other generated by a chatbot. The AI-generated essay should be refined three times and use the same selected cause and effect as the student's essay. The two versions will then be evaluated using an open-sourced digital humanities tool, Voyant Tools, to compare and analyze quantitative and qualitative results. The poster and lightning talk will share key comparative findings and draw conclusions on the linguistic, rhetorical, stylistic, and research choices of the two approaches. This is a work-in-progress looking for constructive developmental feedback.

Bibliography

Zapata, G. C., Saini, A. K., Cope, B., Kalantzis, M., & Searsmith, D. (2024, March 29). Peer-Reviews and AI Feedback Compared: University Students' Preferences. Worktribe.com. https://nottingham-repository.worktribe.com/output/33294270/peer-reviews-and-ai-feedback-compared-university-students-preferences

A Harvard student asked her professors to grade ChatGPT's essays. It got mostly A's and B's. (n.d.). Www.wbur.org. https://www.wbur.org/news/2023/07/26/harvard-student-chat-gpt-experiment-maya-bodnick

Digital Scholarship at Harvard: Current Practices, Opportunities, and Ways Forward. (n.d.). Retrieved September 26, 2024, from https://projects.iq.harvard.edu/files/dsi/files/harvard_ds_final-report_20170627_v2.pdf

Michielin, D. (2023, February 6). Bot or scientist? The controversial use of ChatGPT in science. Foresight. https://www.climateforesight.eu/articles/chatgpt-science/

Picciano, A. G. (2024, June). Graduate Teacher Education Students Use and Evaluate ChatGPT as an Essay-Writing Tool. ResearchGate; The Online Learning Consortium. https://www.researchgate.net/publication/381121445_Graduate_Teacher_Education_Students_Use_and_Evaluate_ChatGPT_as_an_Essay-Writing_Tool

UNESCO, U. (n.d.). Artificial Education in Higher Education (U. UNESCO, Education 2030.) [Review of Artificial Education in Higher Education]. UNESCO. https://library.aup.edu/ld.php?content_id=34684258

Education, J. G. for T. H. (n.d.). British Academics Despair as ChatGPT-Written Essays Swamp Grading Season. Inside Higher Ed. https://www.insidehighered.com/news/global/2024/06/21/academics-dismayed-flood-chatgpt-written-student-essays

Narvika Bovcon[1], Aleš Vaupotič[2]

[1]University of Ljubljana, Faculty of Computer and Information Science, Slovenia; [2]University of Nova Gorica

Did the Avant-Garde Become Materialized in New Media? Techno-Performance and Socialization

Wednesday, 05/Mar/2025 2:00pm - 2:30pm

ID: 266 / Session LP 03: 2

Long paper (full-text) | 20-minute presentation with a 10-minute Q&A

Keywords: New media art, the avant-garde, socialization, infosphere, Nuša Dragan, Srečo Dragan

This contribution builds on Lev Manovich's thesis that avant-garde procedures were incorporated into the very operation of software, as the computer as a meta-medium completely transcodes the reality. The text examines the foundations of computer science in mathematical formalism and considers Turing's computing machine as a starting point for the generative aspect of computer art. The generative art-making was practiced also in the neo-avant-gardes. Data processing introduces into today's world a decoupling between understanding and agency that was once thought to require active thinking. Juri Lotman's idea of periphery-centre dynamic explains the shift of mechanical and (once) avant-garde procedures to the center of the globalized culture. At the same time, the image of a human is radically changing as she finds herself in a new homogeneous environment made of information: she becomes an inforg (Luciano Floridi) together with artificial and hybrid agents. Identity as informational integrity is discussed with the examples of performances and happenings of Srečo Dragan from the conceptualist period, when he collaborated with Nuša Dragan and the extended OHO group. The discussion touches upon the latest new media works by Srečo Dragan, where a multitude of interfaces mediates between the self-awareness (memory, declaration statements) and the body of the participant, and builds an archive of connections for a society established through technology. Dragan’s idea of socialization with art foregrounds a performative educational practice that interferes with the user's identity.

Daniel Brodén[1], Mats Fridlund[1], Leif-Jöran Olsson[2]
[1]Gothenburg Research Infrastructure in Digital Humanities (GRIDH), University of Gothenburg, Sweden; [2]Språkbanken Text, University of Gothenburg, Sweden

The Discourse on Terrorism Vectorised: Mapping the Debate on Political Terror in the Swedish Riksdag, 1968–2021

Thursday, 06/Mar/2025 9:00am - 9:30am
ID: 201 / Session LP 04: 1
Long paper (abstract) | 20-minute presentation with a 10-minute Q&A
Keywords: terrorism discourse, parliamentary data, word vectors, mixed methods

1. Introduction

Political speech is a significant area of interest in the study of the discourse on terrorism. In terrorism studies, considerable critical attention has been devoted to the political discourse on terrorism-terrorism, examining both contemporary and historical perspectives, including the tracing policymaking processes and the ways in which governments and political actors construct narratives around terrorism and counterterrorism (Zulaika 2009). As Strandh and Eklund (2014: 359) notes in their qualitative study of the post-9/11 terrorism discourse in the Swedish Parliament (the Riksdag) (Strandh & Eklund), counterterrorism can be described as a complex and constantly shifting policy area.

1.1 Purpose and Aims

Similar to previous historical research on Swedish parliamentary datasets, this paper focuses on conceptual history (Ohlsson et al. 2022; Brodén et al. 2023; Jarlbrink & Norén 2023), but goes beyond the application of common statistical measurements that has dominated text mining of the debates in the Riksdag (Rouces et al. 2019; Ohlsson et al. 2022; Brodén et al. 2023; Jarlbrink & Norén 2023). Utilising word vectors and a custom-made dataset developed within the Terrorism in Swedish Politics project (2021–2025, SweTerror) (Edlund et al. 2022), the proposed paper maps the transformation of the discourse on terrorism in the Swedish parliamentary debates during the electoral periods 1968–2021. By generating word vectors (Yao et al. 2017), which capture the contextual closeness of words (with semantically similar words represented by numerically close vectors), we analyse and compare the top twenty words related to the base forms of terrorism, segmented by the terms of office of Swedish governments, to trace main periods in the parliamentary discourse on terrorism 1968–2021.

The paper is driven by a straightforward research question: How do the top twenty terrorism-related words reveal distinct shifts in the framing of terrorism in the Swedish Riksdag. This general research question highlights the potential of word vectors to identify patterns of continuity and change in the Swedish parliamentary discourse on terrorism, along with the evolving associations of the term over time. On another level, the paper connects to the broader debate within Digital Humanities regarding the importance of addressing the contextual complexities of text mining large-scale archival collections. As digital historian Jo Guldi (2023) argues, focusing solely on the accuracy and robustness of text mining can only take data-driven analysis so far. Without a contextualised understanding of the materials, the results may raise more questions than they answer (see also Bode 2018). Therefore, we depart from a contextual, mixed methods approach that combines expertise from Language Technology and Digital humanities as well as Digital History and Terrorism Studies.

2. Materials and Methods

We begin by outlining how the SweTerror project combines enriched document annotation and word vectors to trace various aspects of the discourse on terrorism in the Swedish parliamentary debates during the electoral periods 1968–2018. We discuss the customisation of the SWERIK dataset and the use of word vectors as a methodological approach, emphasising the application a contextualising understanding of our parliamentary data.

2.1 The SweTerror Corpus

The SweTerror project utilises an enhanced and curated subcorpus derived from the parliamentary minutes provided by the SWERIK infrastructure (latest version 1.2.0), which structures the Swedish parliamentary records 1867–2023 into a single corpus format (https://github.com/swerik-project/swerik-project.github.io), containing approximately 4 M tokens per parliamentary year (Yrjänäinen et al. 2024). To better align with our research objectives, the SweTerror project has customised the SWERIK dataset, performing additional quality control tasks such as identifying structural issues, since the corpus includes not only 'pure' debate content but also 'secondary' text, such as headings and voting results. (Olsson et al., submitted for publication).

SweTerror aims to enact a contextualised understanding of parliamentary data. By assigning a Persistent Identifier (PID) for the speeches in our dataset, we can leverage SWERIK's metadata on Members of Parliament (MPs) to analyse structural differences in the debates on terrorism at party level. From a mixed methods perspective, our approach emphasises the genre of parliamentary debates and situates the computational results within the dynamics of the Swedish parliamentary discourse, treating the debates as context-bound expressions. While government representatives typically frame their policies positively, opposition debates often focus on reframing the government's narrative rather than reinforcing it. Additionally, our analysis incorporates dynamic time periods (see below) factoring in the parliamentary year (autumn–summer, rather than the calendar year) and electoral periods, the latter being our main focus in this specific paper in order to provide a general scope on the development.

2.2 Word Vectors as Methodological Approach

In the SweTerror project, we combine enriched document annotation with word embeddings to create 'temporal lenses' for analysing the Swedish parliamentary debates during the electoral periods 1968–2018. By using word vectors (Mikolov et al. 2013), we enhance our text mining capabilities, moving beyond simple keyword searches and frequency counts. This approach allows us to identify linguistic patterns and key topics without relying on predefined categories (Olsson et al., submitted for publication). Through iterative experimentation with the parameters for generating usable models from our datasets, we determined that reducing the vector dimensions to 300 provided a good balance between capturing meaningful semantic information and managing the computational resources.

A key aspect of our word2vec pipeline is that the models learn word embeddings through both predicting the surrounding context of a word (skip-gram) and by predicting a word given its surrounding context (Continuous Bag of Words, CBOW). The trained word vectors support various Language Technology tasks, such as tracking changes in language use over time across multiple debate protocols. Notably, the generation of word vectors is performed for all dynamic time periods, meaning the word vectors are recalculated for each distinct period (e.g. parliamentary years and electoral periods). This enables the models to capture shifts in language use and context over time, rather than relying on a single, static set of word vectors, ensuring that temporal changes in meaning and usage are accurately reflected in the analysis.

3. Top Twenty Terrorism Words

We then focus on our use of word vectors to trace the transformation of the discourse on terrorism in the Riksdag. By using similarity word vectors to identify the top twenty terrorism-related words (the terms that semantically behave most similar to 'terrorism') during each electoral period 1968–2018, we can distinguish three distinct periods: 1968–1979, 1980–1997, and 1998–2018. As anticipated, these similarity words include 'core terms' such as 'terrorist', 'terrorist deed' (terroristdåd) and 'terrorist organisation' (terroristorganisation). Furthermore, the top words reflect a focus on policy-making related to counterterrorism, particularly around Sweden's Terrorist Act enacted in 1973, which continued to be a topic of debate (Brodén et al. 2023). However, the top twenty words also indicate a broader shift in the understanding of terrorism, which tentatively can be divided into the three periods below

3.1 Period 1: 1968–1979

The first period reflects the emergence of the modern concept of terrorism, as outlined by Stampnitsky (2013) and Zoller (2021). Previous research has shown that terrorism became a significant political issue in the Swedish Parliament in the early 1970s (Fridlund et al. 2022a; Fridlund et al. 2022b; Brodén et al. 2023), following a series of attacks by militant Croatian exiles linked to the Croatian National Resistance (HNO) and the historical Ustaše movement, including the killing of the Yugoslavian ambassador in Stockholm in 1971 and the armed hijacking at Bulltofta airport in 1972 (Hansén 2007). This was followed by two major incidents associated with the Red Army Faction (RAF): the West German embassy siege in 1975 and a failed kidnapping attempt targeting former Minister Anna-Greta Leijon in 1978. The emergence of the modern notion of terrorism in the Swedish parliamentary debate is evident in the prominence of terms such as 'aircraft hijacking' (flygkapning) and 'hostage' (gisslan) among the top twenty words, reflecting the common tactics of terrorist groups during the 1970s, which centred on political extortion through hostage-taking. Additionally, legislative efforts to curb international terrorism are reflected in the prominence of terms such as 'United Nations' (FN) as well as words such as 'anti-terrorist legislation' (terroristlagstiftning), 'criminal' (brottsling) and 'counter-terrorism' (terroristbekämpning) that are connected to the enactment of the Terrorist Act. The inclusion of terms like 'law of exception' (undantagslag), 'constitution' (konstitution), 'innocent' (oskyldig) and 'rule of law' (rättsstat) further suggests that the debates about the Swedish antiterrorist legislation often focused on criticisms of its nature.

3.2 Period 2: 1979–1997

During 1979–1981, terms such as 'terrorist', 'terrorist act' and 'terrorist organisation' ranked among the top references, with 'hijacking' also making it into the top ten. The Terrorist Act remained a critical issue in the Riksdag, partly due to the controversy surrounding Kurdish nationalists associated with the PKK (Kurdistan Workers' Party) and the unsolved murder of Prime Minister Olof Palme in 1986 (Brodén et al. 2023). From 1982 onwards, words such as 'Kurd', 'Kurdish' and 'PKK' became prominent in the parliamentary discussions on terrorism. However, from 1979 onwards, the debates reflect a broader and less distinct interpretation of terrorism, with the concept increasingly being linked to issues such as the 'narcotics trade' (narkotikahandel) and 'weapons of mass destruction' (massförstörelsevapen) as well as state terrorism, including terms such as 'military dictatorship' (militärdiktatur) and 'military apparatus' (militärapparat). This shift partly corresponds to a heightened focus on the civil wars and insurgencies, driven by Cold War dynamics and U.S. backed efforts to counter perceived communist threats in Latin America. Among the top twenty words during this period were specific references related to state terror such as 'military junta' (militärjunta) and 'junta', reflecting this particular geopolitical context.

3.3 Period 3: 1998–2021

Research has consistently shown the 11 September 2001 attacks in the U.S. to be a watershed moment in the Western discourse on terrorism (for Sweden, see Strandh & Eklund 2014). As 9/11 occurred midway through the electoral period of 1998–2001, we expand our analysis to include these parliamentary years in order to better trace this shift in discourse. Regardless, in the aftermath of 9/11, the top ten words reflect an increased focus on terrorist attacks, with terms such as 'terrorist deed', 'terror attack' (terrorattack) and 'terror threat' (terrorhot) featuring more prominently in the top twenty, indicating a heightened Swedish terrorism-mindedness (Fridlund 2011), that is the domestication of terrorism as a widely recognised security threat within society. Starting around 9/11, terrorism also became increasingly linked to counter-terrorism efforts, as indicated by the emergence of terms like 'countering' (bekämpande) and 'threat image' (hotbild) among the top twenty words. Additionally, terrorism began to be distinctly associated with Islamism. After 2001, Sweden, like many other countries, faced acts of political violence connected to Islamist extremism, including the 2017 truck attack in downtown Stockholm that resulted in five deaths. From 2006 onwards, terms such as 'Islamistic' (islamistisk), 'Islamism' and 'fundamentalism' as well as 'radicalisation' (radikalisering) and 'extremism' became prominent in the discourse.

4. Conclusions

We conclude by synthesising our findings, emphasising how our word vector approach and analytical focus on the top twenty terrorism-related similarity words enable us to map the transformations of the discourse on terrorism in the Riksdag at a high level. To further state our case, we also highlight structural differences in word usage at the party level, including distinctions between government and opposition during different electoral periods.

Bibliography

Bode, Katherine (2018): A world of fiction. Digital collections and the future of literary history, Ann Arbor, MI: University of Michigan Press.

Brodén, Daniel, Mats Fridlund, Leif-Jöran Olsson, Magnus P. Ängsal, Patrik Öhberg (2023): 'The Diachrony of the New Political Terrorism: Tracing Neologisms and Frequencies of Terror-related Terms in Swedish Parliamentary Data 1971–2018', DHNB 2022: Proceedings, CEUR-WS.

Ditrych, Ondrej (2014): Tracing the discourses of terrorism, Palgrave Macmillan.

Edlund, Jens, Daniel Brodén, Mats Fridlund, Cecilia Lindhé, Leif-Jöran Olsson, Magnus P. Ängsal, Patrik Öhberg (2022): 'A multimodal digital humanities study of terrorism in Swedish politics: an interdisciplinary mixed methods project on the configuration of terrorism in parliamentary debates, legislation, and policy networks 1968–2018', In Kohei Arai (ed), Intelligent Systems and Applications: IntelliSys 2021. Lecture Notes in Networks and Systems (Vol. 295). Cham: Springer.

Fridlund, Mats (2011): 'Bollards, buckets and bombs', History and technology, 27:4.

Fridlund, Mats, Daniel Brodén, Leif-Jöran Olsson & Magnus P. Ängsal (2022a): 'Codifying the debates of the Riksdag: Towards a framework for semi-automatic annotation of Swedish parliamentary discourse', In Matti La Mela, Fredrik Norén & Eero Hyvönen (Eds.): Proceedings of Digital Parliamentary Data in Action (DiPaDa 2022). Workshop Co-located with the 6th Digital Humanities in the Nordic and Baltic Countries Conference (DHNB 2022), Uppsala, Sweden, March 15, 2022, CEUR Workshop Proceedings.

Fridlund, Mats, Daniel Brodén, Victor Wåhlstrand Sjöström (2022b): 'The diachrony of political terror: Tracing terror and terrorism in Swedish parliamentary data 1867–1970', DHNB2023 Conference Proceedings, 5:1, 79–89.

Guldi, Jo (2023): The dangerous art of text mining: A methodology for digital history, Cambridge: Cambridge University Press.
Hansén, Dan (2007): Crisis and perspectives on policy change, diss., FHS.

Jarlbrink, Johan & Fredrik Norén (2023): 'The rise and fall of "propaganda" as a positive concept: a digital reading of Swedish parliamentary records, 1867–2019, Scandinavian Journal of History, 48:3.

Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg S Corrado, Jeff Dean (2013): 'Distributed representations of words and phrases and their compositionality', Advances in Neural Information Processing Systems. arXiv:1310.4546. Bibcode:2013arXiv1310.4546M.

Ohlsson, Claes, Viktor Wåhlstrand Skärström, Henrik Björck (2022): 'The market as a concept in Swedish parliamentary records from 1867 to 1970: A mixed methods study'. In Digital Parliamentary Data in Action (DiPaDA 2022) workshop, Uppsala University, Sweden, March 15, 2022, CEUR-WS.

Olsson, L-J, D Brodén, M P Ängsal, M Fridlund, P Öhberg (submitted for publication): 'Augmented Analysis of Parliamentary Debates: The Word Embedding and Context-sensitive Approach of the SweTerror Project', Proceedings of the Digital Parliamentary Data in Action (DiPaDA 2022) Workshop co-located with 8th Digital Humanities in the Nordic and Baltic Countries Conference (DHNB 2024), Reykavik, Iceland.

Rouces, Jacobo, Lars Borin, Nina Tahmasebi (2019): 'Political stance analysis using Swedish parliamentary data', In Workshop Proceedings (Vol. 2364). Digital Humanities in the Nordic Countries 4th Conference. Aachen: CEUR Workshop Proceedings.
Stampnitzky, Lisa (2013): Discipling terror, Cambridge University Press: Cambridge.

Strandh, Veronica, Niklas Eklund (2014): 'Swedish Counterterrorism Policy: An Intersection Between Prevention and Mitigation?', Studies in Conflict and Terrorism, 38:5, 359–379.

Yao, Zijun, Yifan Sun, Weicong Ding, Nikhil Rao, Hui Xiong (2017): 'Dynamic word embeddings for evolving semantic discovery', International conference on web search and data mining, WSDM.

Yrjänäinen, Väinö, Fredrik Mohammadi Norén, Robert Borges, Johan Jarlbrink, Lotta Åberg Brorsson, Anders P Olsson, Pelle Snickars, Måns Magnusson (2024): 'The Swedish Parliament Corpus 1867–2022', LREC-COLING 2024: 16100–16112.

Zoller, Silke (2021): To deter and punish: Global collaboration against terrorism in the 1970s, Columbia University press.

Zulaika, Joseba (2009): Terrorism: The self-fulfilling prophecy, University of Chicago Press.

Jørgen Burchardt
Middelfart Museum, Denmark

Analytical Evaluation of OCR Accuracy in Historical Newspaper Advertisements: Challenges and Solutions

Thursday, 06/Mar/2025 2:00pm - 2:20pm
ID: 138 / Session SP 08: 1
Short paper (abstract) | 15-minute presentation with a 5-minute Q&A

Keywords: digital newspaper archive, OCR, source criticism, dataset

Newspaper advertisements provide valuable insights into historical trends and societal changes. OCR (Optical Character Recognition) created databases of digitized newspapers offer vast datasets that allow researchers to explore shifts in employment patterns, consumer behavior, and industry evolution. For example, these databases can help answer questions like when office delivery jobs vanished or when men replaced women in dairies.

However, OCR accuracy when processing advertisements can be inconsistent, raising concerns about the reliability of data and the conclusions drawn from it. This study examines whether OCR errors in advertisements are significant enough to affect findings after applying strict source-critical analysis. Historians, who already face challenges in scrutinizing sources, now must also evaluate the reliability of digital archives. This article explores these challenges, revealing that OCR error rates, combined with biases from word-frequency datasets, can compromise the accuracy of historical research due to issues in the shaping of the OCR corpus and later post-processing.

The study involves an empirical examination of OCR outputs from newspapers across different time periods, different countries, and created by different digitization technologies. Previous research (Burchardt, 2024) highlighted major inaccuracies in some archives, with significant OCR error rates. For historians striving for near-perfect accuracy, these rates pose problems, especially in establishing "first-time" events or tracking precise historical trends.

A prior study (Burchardt, 2023, p. 43) indicated that OCR error rates in newspaper advertisements could reach up to 30%, even when body text in the same publications has only an error rate on 10 %. Announcements alone could have error rates as high as 15% even when the body text was nearly flawless, further complicating the analysis of commercial and public messaging. This proposed theory will be examined further.

It is well known that there has been a progress of OCR accuracy as the art of printing improved. Better physical quality of the type pieces improved the accuracy as better designs of the type itself gave a better recognition. These improvements have been neither uniform nor continuous. Early newspaper samples show that OCR accuracy progressed as paper quality and printing techniques advanced. However, between 1730 and 1830, error rates unexpectedly rose, despite general improvements in print quality. This can be attributed to the shift to broadsheet formats, which introduced multi-column layouts that complicated OCR recognition, particularly in advertisements.

Another spike in error rates occurred around 1880, doubling until it gradually declined around 1910. The complexity of advertisement layouts, distinctive fonts, and the use of images likely contributed to this trend.

The research will be carried out in two phases. The first phase will involve small-scale tests across different time periods since 1700, newspapers printed with Latin letter types, and archives to identify high-error periods or formats for further investigation in phase two. The newspapers are chosen from digital archives where the pure OCR text can be viewed. Errors are identified by manually reading the OCR text and comparing it with the graphic image of the newspaper page.

In the second phase, a more focused analysis will be conducted on samples from the periods, newspapers, or archives identified as having the highest error rates. The errors will be thoroughly examined, and their causes identified. The result should be a base for a discussion of solutions through re-OCRing or post-processing methods.

Bibliography

Burchardt, Jørgen: "Source criticism, bias, and representativeness in the digital age: A case study of digitized newspaper archives" in DHNB2024 From Experimentation to Experience: Lessons Learned from the Intersections between Digital Humanities and Cultural Heritage May 29 – 31, 2024, Reykjavik, Iceland. Digital Humanities in the Nordic and Baltic Countries Publications

Burchardt, Jørgen: "Are searches in OCR-generated archives trustworthy?" in Jahrbuch für Wirtschaftsgeschichte 2023; 64(1), p. 31-54.

Guillem Castańar Rubio[1], **Anastasiya Fiadotava**[2], **Agata Hołobut**[3], **Jan Rybicki**[3]
[1]Tallinn University; [2]Estonian Literary Museum; [3]Uniwersytet Jagielloński, Poland

ID: 195 / Poster Session 2: 27
Long paper (abstract) | 20-minute presentation with a 10-minute Q&A
Keywords: Online Humour, Correspondence Analysis, Principal Components Analysis

Testing Quantitative Methods on Multinational Humorous Data

Introduction

The presentation will use the data that were collected within the framework of the CELSA network project "Humour and Conflict in the Public Sphere: An interdisciplinary analysis of humour controversies and contested freedoms in contemporary Europe" (Chłopicki et al., 2024). The project focused on humour that revolves around controversial events in four countries – Estonia, Belarus, Poland, and Belgium.

Material

In each of the countries we have identified two or three **events** (9 in total, one of them yielding data in three countries) that provoked plenty of humour in the public spheres of these countries in 2022–2024. We collected 50 **humorous items** per event (550 items in total). The selection criteria for the 50 items per country that were included in the final dataset were aimed at ensuring their diversity in terms of (a) sources of data, (b) genres and formats, (c) specific topics represented within the broader topic of a controversial event, (d) professional or amateur creators of the humorous content. However, despite our intention to make our dataset as diverse as possible, some of the topics and genres were more prominently represented in our dataset due to their general popularity in the initial full datasets.

The data were collected manually by (a) browsing the most popular humorous media outlets (such as jokes and memes aggregators, social media groups and profiles dedicated to humour, satirical news portals etc.) and (b) searching by the keywords related to the controversial events in relevant languages in the mainstream and social media of the four countries.

The humorous items were coded via the multiple choice Qualtrics survey according to several categories that were jointly developed by the participants of the above-mentioned CELSA network project:

- Genre or combination of genres (video recording of an event, photo, cartoon, internet meme (image only), internet meme (image and text), internet meme (video), text-only joke, humorous comment, satirical news article, blog post, stand-up performance, television comedy, non-humorous comment, or other);

- Presence of verbal, visual, or both verbal and visual elements;

- Humour mechanisms (humorous stereotype, sexual innuendo, status reversal or challenging, transgression, grotesque, juxtaposition of text and image, parody, caricature, ambiguity, exaggeration, irony, recontextualization, word play);

- Communication style of the items that had verbal element(s) (direct and/or based on overstatement, or indirect and/or based on understatement) (Gudykunst and Ting-Toomey, 1988);

- Rhetorical format of the items that had verbal element(s) (statements, questions, commands/imperatives, verbless phrases, expletives, paraverbal comments, longer texts including several of the above).

We have also selected 25 most commented items out of 50 per event and coded the **comments** to those items. The comment sections of the humorous items that we analysed were open for anyone to comment; in social media such as Facebook, internet users commented using their personal accounts while the comments on news media (including the satirical news portal) were anonymous. We looked at the humorousness of the comments (distinguishing between humorous comments, non-humorous comments, unlaughter and unclear comments) and their assessment of humorous items (positive/negative/neutral or unclear). We also looked at the format of the comments and categorised them in the following way: verbal comments only, emoticons only, comments in the form of images/gifs, combination of verbal and non-verbal comments and hyperlinks. For the comments that had verbal elements, we also coded their rhetorical format (see categorisation of rhetorical formats above). Finally, we calculated the number of meta-comments, comments about the people who posted the items / other users in the comment thread and meta-linguistic comments.

The aim of this quantitative analysis consisted in finding associations between the various variables, focusing largely on the associations between particular events and aspects of humorous items, and their related comments.. Numerical values were assessed with standard principal components analysis and categorical values were submitted to correspondence analysis.

Results

Correspondence analysis of categorical values over the entire set of categorical variables (country, genre, visual/verbal, humour mechanism, communication style, and rhetorical format) grouped strongly by country of origin (Fig. 1). In other words, humorous items created in response to the events included in the study were quite different depending on whether the humour creators and sharers were from Belarus, Belgium, Estonia or Poland. In fact, any greater overlap there only happened between Belarusian and Estonian humorous items. This separation prevailed in a majority of combinations of variables; in another instance (Fig. 2), the country of origin seemed to determine the various humour mechanisms applied.

Figure 1. Correspondence analysis of all categories in all humorous items that responded to events; colours denote the humorous items' country of origin.

[Fig. 1 goes here]

Figure 2. Correspondence analysis of three humour mechanism variables in all humorous items; colours denote the humorous items' country of origin.

[Fig. 2 goes here]

In an attempt to look for differences between humorous items that responded to particular events rather than between the countries of origin, numerical results associated with comments to humorous items that responded to a particular event were pooled, producing 8 single-country events and a single 3-country (Belarusian, Estonian, Polish) event (Prigozhin's coup). This time, comments were classified into humorous, non-humorous, unlaughter ("this is something not to be laughed at", see Billig, 2005, p. 192) and unclear. This is presented in a barplot in Fig. 3: note the differences between comments to humour targeting

Prigozhin's coup in Belarus and Poland (a relative balance between humorous and non-humorous) and the predominantly humorous comments in Estonia.

A similar analysis was conducted in the same way for commenter's assessment (as positive or negative) of the humorous items (Fig. 4). Here, the results were much less clear; visibly positive comments were the highest in two cases of humour targeting living politicians: Kaczynski in Poland and Kallas in Estonia.

As in both of the above cases the differences shown proved to be of no statistical significance, principal components analysis was used to better assess comments to the humorous items related to particular events. In terms of humorous/non-humorous etc., Polish and Belarusian comments were both quite compact, and Estonian ones differed the most between themselves (Fig.5).

In terms of positivity, the Polish Daukszewicz case was a clear outsider (Fig. 6) because the humour related to this case inspired more negative comments than the humour related to the other cases that we have analysed. Once again, Belarusian and Estonian comments to humorous items were quite similar and, at the same time, removed somewhat from those in Poland. Within same-country comment pools, the greatest variety was observed in those in Polish; the greatest similarity was that between the two Belgian sets. In both PCA analyses, the two principal components (PC1 and PC2) explained almost 100% of variance, which makes these results reliable.

Figure 3. Humorous, non-humorous, unlaughter and unclear comments on humorous items that responded to particular events (with separate country categories for Prigozhin's coup).

[Fig. 3 goes here]

Figure 4. Positive, negative and unclear comments on humorous items that responded to particular events (with separate country categories for Prigozhin's coup).

[Fig. 4 goes here]

Figure 5. PCA analysis of humorous, non-humorous, unlaughter and unclear comments on humorous items related to particular events (with separate country categories for Prigozhin's coup).

[Fig. 5 goes here]

Figure 6. PCA analysis of positive/negative comments to the humorous items related to particular events (with separate country categories for Prigozhin's coup).

[Fig. 6 goes here]

Conclusions

Correspondence analysis of the variables pertaining to the humorous items and comments to them showed that many of these variables grouped around the country of origin of the humorous items. Therefore, despite the globalisation of (humorous) online communication, national differences still do play a role in humour production. On the humour reception end (i.e. comments to the humorous items) the results were slightly more homogeneous in terms of both humorousness/non-humorousness and positivity/negativity. However, PCA showed that the differences in humour reception can sometimes stem from the differences between particular cases, and also the differences between the countries of origin within a single case, as in the three national variations in the case of Prigozhin's coup.

Bibliography

Billig, M. (2005). Laughter and ridicule: Towards a social critique of humour. Thousand Oaks: Sage.

Chłopicki W., Kuipers G., Laineste L., Castañar G., Fiadotava A., Hołobut A. and Nicolaï J. (2024). Specific Humor Scandal Database. KU Leuven RDR, https://doi.org/10.48804/PTPQVB

Gudykunst W.B. and Ting-Toomey S. (1988). Culture and Affective Communication. American Behavioral Scientist, 31(3), 384–400. https://doi.org/10.1177/000276488031003009

Coppélie Cocq[1], **Koraljka Golub**[2], **Marianne Gullberg**[3], **Mats Fridlund**[4]
[1]Umeå University, Sweden; [2]Linné University; [3]Lund University; [4]University of Gothenburg

Huminfra – A Swedish national research infrastructure for digital and experimental humanities

ID: 120 / **Poster Session 2: 21**
Poster and demo (abstract) with accompanying a 1-minute lightning talk
Keywords: Infrastructure, Experimental Humanities, Digital Humanities

Huminfra is a Swedish national research infrastructure supporting digital and experimental research in the Humanities. It is a consortium consisting of 12 nodes across 11 universities and organisations, coordinated by Lund University Humanities Lab and is funded by the Swedish Research Council and the consortium nodes.

Huminfra acts in three domains.

(1) Huminfra provides users with a single entry point for finding existing Swedish materials, research tools, and experts. The web-based information platform www.huminfra.se compiles this information and links to Swedish digital/e-scientific data sets, tools, expertise and educational opportunities and makes these easily identifiable and accessible. Huminfra.se contributes to a more efficient use of national resources at the cutting edge.

(2) Huminfra creates and holds new national courses in e-scientific, digital, and experimental methods drawing on the expertise at its nodes.

(3) As of July 2024, Huminfra hosts DARIAH-SE, acting as a national node in the European Research Infrastructure Consortium, *The Digital Research Infrastructure for the Arts and Humanities* (DARIAH ERIC).

Huminfra's vision is to strengthen and promote Swedish Humanities and ensure its international competitive edge. Huminfra will reinforce research in areas of specific interest to Sweden (cultural heritage, languages spoken in Sweden, national archives) and make such research and resources internationally visible and accessible through DARIAH-SE. Huminfra will enhance the visibility of the Humanities, create new possibilities for innovation and impact, and promote collaboration with societal stakeholders such as cultural institutions, education, health and industry.

This poster will include examples of activities conducted and planned by the Huminfra consortium, and examplify how this infrastructure can contribute to advancing the digital humanities in Sweden as well as in the Nordic and Baltic countries.

Coppélie Cocq, **Stefan Gelfgren**, **Lars Samuelsson**, **Jesper Enbom**
Umeå University, Sweden

Keeping pace with digital transformation: the case of senior citizens in Sweden

Wednesday, 05/Mar/2025 1:30pm - 2:00pm
ID: 163 / Session LP 02: 1
Long paper (abstract) | 20-minute presentation with a 10-minute Q&A
Keywords: digital divide, senior citizens, digital transformation

The access and processing of digital data are expected to transform society, as oil did in the 19th century (Coultry & Mejias 2019), and lead to new innovations, sustainable business models, and a more equal society. A fundamental part of this transformation is the flow of data between different actors (within business, security and welfare, etc.) for different purposes. In the digital transformation process, previously analogous data are digitized, and new data emerge, that are collected, coordinated and processed, in order to develop new allegedly efficient services.

This has been a process for years, and today many digital services are intertwined with everyday life. In many cases, life gets easier through easy access to digital services – for communication with other individuals (such as family and friends) and authorities (for example welfare institutions, societal institutions, and banks). However, in this process, some people benefit, and some have, for various reasons, difficulties in keeping up the pace with the digital transformation, and are consequently left behind, which is often referred to as an emerging "digital divide" (Van Deursen & Van Dijk 2014; Van Dijk & Hacker 2003). During the last decade there has been emerging research on the existence of a "grey divide" (Friemel 2014, Quan-Hase 2018) where older people are excluded from the digital transformation. This paper discusses to what extent and how such dividing processes affect senior citizens.

More specifically, in this paper we will present a case-study that investigates how senior citizens in Sweden make sense of digital transformation, i.e. what limitations and strategies they experience in their everyday life. It is based on 6 focus group interviews with a total of 29 participants between the age of 70 and 85.

According to the latest report about The Swedes and the internet (Svenskarna och internet 2024), around 50 % among Swedes 74 years and older use internet on a daily basis while approximately 20 % use it more seldomly. In this group around 30 % declares themselves as non-users. It is also in the oldest group, and especially among women, the self-reported need for digital help is the highest. It is six times bigger compared to the need among Swedes in working age. Sweden is also of special interest in digitalization studies as it stands out as an exception on a global scale. Swedes show high levels of trust in authorities and fellow citizens (see, e.g., Delhey & Newton 2005) and have a high degree of internet connectivity (among the top 10 globally, see DataReportal 2019). Moreover, they tend to value self-fulfilment while they show low levels of affiliation to traditional structures (Sweden being on the top right corner in the Inglehart & Welzel's cultural map of the world, 2023). Hence, a Swedish case can indicate what the future of digital transformation in a global context may look like.

Our research group examines the tension between the societal process of digital transformation, pushed by political and business agendas, and the importance of privacy and individual integrity. In Sweden, contacts with banks, insurance agencies, and authorities are made through online services, with a digital identification system at the center of it all – a system issued by the banks. Activities such as buying a parking ticket, booking and buying travels or events, and making reservations at restaurants, take place online to a great and increasing extent. At the same time, cash is almost not used any more, and for many types of purchases (such as buying bus and parking tickets, but also purchases in many regular stores), it has almost become standard that cash is not accepted. Instead, card or a digital service for transferring money is the go-to solution. In other words, digital services are deeply intertwined with being a citizen.

The interviews addressed the topics of everyday use of the internet and digital devices, advantages and disadvantages with digital technologies, reflections about data collection and digital transformation, and about situations and circumstances that could or could not be seen as legitimizing data collection.

A thematic analysis resulted in identifying the predominant themes of anxiety toward the process of digital transformation and digital technologies; the social aspects of these technologies; the experienced lack of control in this process (and various feelings associated to this); and the ambiguity of on the one hand being forced into a process, and on the other experiencing new possibilities.

Our results and analysis indicate that for senior citizens, the questions of data collection and personal privacy are less important issues than practical everyday concerns. They describe how digital transformation impacts on everyday situations and discuss both obstacles and solutions. A recurring theme is the imposed limitations of digitalization. Rather than facilitating everyday life, in many cases the transition towards digital solutions is experienced as aggravating. For instance, several interviewees describe how they or their friends refrain from taking the car downtown, or travelling by bus, due to how these activities would force them to use digital applications they do not master. At the same time, they see the benefits of digital devices in other areas – a majority use them to keep in touch with friends and family (through social media and communication apps), to read the news, to watch films, and listen to books, etc. Furthermore, many of the interviewees express concern for society in general (rather than for themselves), and for the future. Relying on digital solutions is perceived as leading to a more vulnerable society.

Finally, in light of these concerns, the paper discusses the role of various actors in society in providing support for senior citizens for keeping pace with digital transformation in Sweden. Typically, the experience of the senior citizens interviewed is that many of the actors who impose digital "services" – who enforce the use of digital devices and applications – like banks, commuting and parking companies, and actors in the care sector, do not provide the support needed to comfortably use these services, leaving the user in a state of insecurity and often a feeling of insufficiency. The support is rather to be found among friends and family, by turning to the municipal library, or by being lucky enough to find an engaged person to ask for help. This, in turn, may lead to feelings of vulnerability, shame over one's experienced shortcomings, and guilt over "using" friends and family.

Investigating how senior citizens navigate the increasing digitalization of society clearly discloses the extreme pace with which this transformation takes place – a pace that it is difficult even for society at large to keep up with. In particular, it exposes many of the challenges that accompany digitalization – it highlights the digital divide, and it shows the need for society (and other actors) to take responsibility for not leaving groups of people behind. At the same time, it reveals many of the ways in which digital services and devices can benefit people. Even though our interviewees expressed worries about the current development and how digitalization enforces unwanted changes in their everyday lives, they also emphasized its positive sides; ways in which it facilitates their lives and is used to increase wellbeing.

Bibliography

Couldry, N., & Mejias, U. A. (2019). The costs of connection: how data is colonizing human life and appropriating it for capitalism (1st ed.). Stanford University Press. https://doi.org/10.1515/9781503609754

Friemel, T. N. (2016). The digital divide has grown old: Determinants of a digital divide among seniors. New Media & Society, 18(2), 313–331. https://doi.org/10.1177/1461444814538648

Quan-Haase, A., Williams, C., Kicevski, M., Elueze, I., & Wellman, B. (2018). Dividing the Grey Divide: Deconstructing Myths About Older Adults' Online Activities, Skills, and Attitudes. The American Behavioral Scientist (Beverly Hills), 62(9), 1207–1228. https://doi.org/10.1177/0002764218777572

Van Deursen, A., & Van Dijk, J. (2014). The digital divide shifts to differences in usage. New Media & Society, 16(3), 507–526.

Van Dijk J and Hacker K (2003) The digital divide as a complex and dynamic phenomenon. Information Society 19(4): 315–326.

World Values Survey (2023). Inglehart & Welzel's cultural map of the world. https://www.worldvaluessurvey.org/WVSContents.jsp?CMSID=Findings

Trausti Dagsson
The Árni Magnússon Institute for Icelandic Studies, Iceland

Digitizing the Folksong Collection of Bjarni Þorsteinsson: Musical Notations as Data in a Wider Context

Thursday, 06/Mar/2025 2:20pm - 2:40pm
ID: 196 / Session SP 05: 2
Short paper (abstract) | 15-minute presentation with a 5-minute Q&A
Keywords: music, notation, folkmusic, digital archive

The Icelandic folksong collection of Bjarni Þorsteinsson (1861-1938) was a ground-breaking work when published in the years 1906 – 1909. The content of the collection is divided into two parts; songs found in various manuscripts and previously published sources, and songs collected from oral history from a number of informants. In the collection, information about each song can be found along with information about the informants, known origin of the song, different variants from other informants and most importantly the notation of each song. Bjarni Þorsteinsson was a priest in northern Iceland, educated in music and had a passion for preserving older folk music in a time where newer songs were getting more popular with the risk of the older songs being forgotten. Using similar methods as other folklore collectors at the time, for example the folklore Jón Árnason, he collected both by himself and got material from colleagues who collected for him.

The collection, republished as one book in 1974 has been digitized by the National Library of Iceland on the website Baekur.is (e. books.is) which gives access to digital reproductions of old Icelandic books with the aim of enabling access to all published books prior to 1870. However, when it comes to notations, that digital version is limited to only images of the scanned pages.

This abstract presents a pilot project where notations were made digital using musical score software, a work that was carried out as a summer job for a student in music with a background in folk music. In this pilot project we only focused on the part of the collection that consists of songs collected from oral history since it gives ways of linking it to the etymology collection of the Árni Magnússon Institute for Icelandic Studies which consists of audio recordings, many of which include singing or playing of folksongs. The result of this work, around 600 digital notations, has been made available at the website Ismus.is which is a database of Icelandic folk tradition and musicology. The website Ísmús contains for example Sagnagrunnur, a database of printed Icelandic folk legends, digital access to the previously mentioned audio archive and a database of folk poetry among other data.

The result of this project has been made available as part of the website. Users can listen to the songs in the digitized folk song collection, download them as MusicXML files and search for songs using a simple melody-based search engine. The database also links informants and collectors of music to a wider context since many of these people were also informants in other relevant collections, for example informants of folk legends.

In the presentation, the digital methods used in this project will be discussed. It will be demonstrated how notations have been thought of as data and converted to a system of coding that approaches the music as simple data with the possibility of searching for a part of melody and similar melodies. Experiments in the early stages on how to link the notations to sound data in the audio archive will also be explored. Our work will also be compared to other similar projects, for example the music part of the collection Finlands svenska folkdiktning which has been made available digital by Svenska Litteratursällskaped i Finland (Swedish Litteratus Society of Finland) and the Dutch Song Database made by Meertens Institute in the Netherlands.

Jānis Daugavietis, Sanita Reinsone
Institute of Literature, Folklore and Art - University of Latvia, Latvia

Challenges in Assessing the Ecosystem of DH Tools and Resources in Latvia

Wednesday, 05/Mar/2025 4:00pm - 4:20pm
ID: 227 / Session SP 01: 1
Short paper (abstract) | 15-minute presentation with a 5-minute Q&A
Keywords: Latvian DH ecosystem; inventory of DH tools and resources; DH policy; open science

In recent years, the term “DH ecosystem” has been used more and more frequently, usually referring to the national level (Pawlicka-Deger 2022). With the strengthening and expansion of EU research consortia such as CLARIN and DARIAH, there is a need to move from the level of analysis and improvement of specific DH tools and resources to the level of not only DH infrastructure, but even the whole ecosystem. Also from a national perspective, it is essential to map the local DH ecosystem in order to avoid duplication of functions and to ensure a more rational use of public funding, as most DH tools and resources are developed and maintained with public money. This is particularly important in small and relatively poor countries with shrinking populations such as Latvia.

The Latvian DH ecosystem started to emerge in the late 1950s (Skadiņa 2021). In the last 15-20 years, the number of DH tools and resources has grown particularly intensively in various fields of humanities (Daugavietis et al. 2022), including linguistics and language technologies (Grasmanis et al. 2023, Paikens et al. 2023, Saulīte et al. 2022, Saulīte&Darģis et al. 2022, Juško-Štekele&Kļavinska 2022), literary studies (Eglāja-Kristsone 2022), history, cultural heritage, folklore and other disciplines (Apenīte et al. 2022, Ernštreits 2019, Reinsone & Laime 2022). Most of the DH tools and resources used in the academy and citizen science are developed by public institutions (universities, libraries, archives, museums). The Republic of Latvia has a unitary and centralised state structure, but this is not the case for DH tools and resources, even though they are built and maintained with public funding (including EU). They are created and maintained by relatively independent institutions belonging to different fields (research, heritage conservation), e.g. university institutes. At the same time, science policy makers have an interest in ensuring that the infrastructures and specific tools and resources in the Latvian DH ecosystem do not duplicate functions and that the field as a whole is not unnecessarily fragmented. It would also be important to understand the current open science practices in the ecosystem of digital humanities resources and tools.

We are conducting research for the Ministry of Education of the Republic of Latvia in the project “Towards Development of Open and FAIR Digital Humanities Ecosystem in Latvia (DHELI)”, where one of the tasks is to map the Latvian DH ecosystem and inventory specific DH tools and resources. Conceptually, this means defining the current ecosystem of digital humanities resources and tools, which will be used and further documented in the future DARIAH-LV framework. In practice, this means creating an inventory tool that will be used to test each of the ~100 tools and resources included in our 'Latvian DH Tools and Resources Database'.

In this paper we will outline the main points of discussion and potential problems in working on this work. From the challenges in creating and classifying the Latvian DH tools and resources sample database (as the same evaluation criteria cannot be applied to different tools and resources), to the technical challenges of the inventory. A separate but very important issue is the communication of the inventory results with the study client and the developers and maintainers of DH tools and resources themselves.

Bibliography

Apenīte, M., Bojārs, U., Garda, A., Goldberga, A., Kreislere, M., Rašmane, A., & Ungure, E. (2022). LNB atvērtie dati pētniecībai (Open Research Data of the National Library of Latvia). Letonica, 47, 168–186.

Daugavietis, J., Karlsone, A., Kunda, I., & Kristāla, A. (2022). Tools and resource development practices in Latvian digital humanities. Letonica, 47, 12–51.

Eglāja-Kristsone, E. (2022). Digital resource Literatura.lv: An introduction. Colloquia, 47, 149–159.

Ernštreits, V. (2019). Electronical resources for Livonian. Proceedings of the Fifth International Workshop on Computational Linguistics for Uralic Languages, 184–191.

Grasmanis, M., Paikens, P., Pretkalnina, L., Rituma, L., Strankale, L., Znotins, A., & Gruzitis, N. (2023). Tēzaurs.lv – The experience of building a multifunctional lexical resource. In Electronic lexicography in the 21st century (eLex): Invisible lexicography.

Juško-Štekele, A., & Kļavinska, A. (2022). Mūsdienu latgaliešu valodas runas korpusa izveide mazāk lietoto valodu dokumentēšanas kontekstā (Creation of Contemporary Latgalian Speech Corpus in the Context of Documenting Lesser Used Languages). Letonica, 47, 226–241.

Paikens, P., Klints, A., Lokmane, I., Pretkalnina, L., Rituma, L., Stade, M., & Strankale, L. (2023). Latvian WordNet. In Proceedings of the 12th Global Wordnet Conference. Global Wordnet Association.

Pawlicka-Deger, U. (2022). Infrastructuring digital humanities: On relational infrastructure and global reconfiguration of the field. Digital Scholarship in the Humanities, 37(2), 534-550.

Reinsone, S., & Laime, S. (2022). Latviešu folkloras krātuves digitālais arhīvs garamantas.lv: izveidošana un attīstības perspektīvas (Digital Archive of Latvian Folklore Garamantas.lv: Development perspectives). Letonica, 47, 50–69.

Saulīte, B., Nespore-Berzkalne, G., Rituma, L., Lasmanis, V., & Gruzitis, N. (2022). Latviešu valodas FrameNet korpuss. Letonica, 47, 284–296.

Saulīte, B., Darģis, R., Gruzitis, N., Auzina, I., Levāne-Petrova, K., Pretkalniņa, L., Rituma, L., Paikens, P., Znotins, A., Strankale, L., Pokratniece, K., Poikāns, I., Barzdins, G., Skadiņa, I., Baklāne, A., Saulespurēns, V., & Ziediņš, J. (2022). Latvian National Corpora Collection – Korpuss.lv. Proceedings of the Thirteenth Language Resources and Evaluation Conference, 5123–5129.

Skadiņa, I. (2021). 'Datorlingvistika Latvijā', Nacionālā Enciklopēdija (Latvijas Nacionālā bibliotēka) <https://enciklopedija.lv/skirklis/106524>;

Vilius Dranseika
Jagiellonian University, Poland

Delineating Philosophy of Medicine: A Data-Driven Approach

Friday, 07/Mar/2025 2:30pm - 2:50pm
Warning: The presentations finish prior to the end of the session!
ID: 107 / Session LP/SP 09: 4
Short paper (abstract) | 15-minute presentation with a 5-minute Q&A
Keywords: philosophy of medicine, bioethics, topic modelling, citation analysis, humanities

Recent discussions tend to describe the philosophy of medicine as (a) a branch of the philosophy of science and (b) a field distinct from medical ethics (e.g., Gifford 2011; Reiss & Ankeny 2022; Schramme 2017). For instance, Thompson & Upshur write: “We treat philosophy of medicine as a branch of philosophy of science. Consequently, ethics does not play a large role” (2018: 5). Others, while conceding that medical ethics is a branch of the philosophy of medicine, still insist that it is possible to discuss the philosophy of medicine without engaging with medical ethics (e.g., Stagenga 2018). If these characterizations are accurate, there should exist various observable patterns (citation patterns, publishing patterns, patterns in the topical composition of papers, patterns in the self-identification of practitioners, etc.) that could be studied by triangulating various metascience approaches.

In this paper, we attempt to delineate and characterize the philosophy of medicine in a data-driven way. Most centrally, our question is whether there is a detectable and characterizable distinction between the philosophy of medicine and bioethics. While our primary aim is to contribute to the understanding of the philosophy of medicine, we also hope that our approach can be used more broadly to study how academic disciplines — including those within the humanities, as is the case in the present study — relate to their neighboring fields. To achieve this, we draw on several data sources (e.g., a full-text corpus of approximately twenty thousand articles from seven leading journals in philosophy of medicine and bioethics, Web of Science incoming and outgoing citation data for these articles, self-selected keywords scholars use to describe their areas of interest on Google Scholar) and apply several analytic approaches (e.g., topic modeling, community detection algorithms, citation analyses).

Triangulating several different metascience approaches produces the following picture: While the data-driven characterization of the philosophy of medicine reliably captures classical issues associated with a relatively narrow understanding of the philosophy of science (e.g., discussions on causality and explanation in medicine, the epistemology of medical diagnosis, concepts of disease and health), it also captures topics that require a broader notion of the philosophy of science (e.g., phenomenological and biopolitical reflections on medicine, the role of religion in medical practice). Furthermore, while the philosophy of medicine seems distinguishable from medical ethics along several parameters (from association with different sets of journals to differences in the self-identification of practitioners), this distinction is somewhat problematic due to several factors. First, metaethical reflections on bioethics (most notably the principlism debate) are firmly rooted in the philosophy of medicine, prompting a qualification of the claims that “ethics does not play a large role” in the philosophy of medicine. Second, philosophical discussions on beginning-of-life issues (most notably debates on the metaphysical and normative status of embryos) exhibit patterns that do not align neatly with a clear distinction between the philosophy of medicine and bioethics. While these debates have some properties strongly associated with the philosophy of medicine (e.g., a relatively high rate of citations to philosophy journals and framing in terms of conceptual analysis), they also have properties unusual for the philosophy of medicine (e.g., they are more prominent in bioethics than in philosophy of medicine journals).

Bibliography

Gifford, F. (2011). Philosophy of medicine: Introduction. In. Gifford, F. (Ed.). Philosophy of Medicine. North-Holland, pp. 1-12.
Reiss, J., & Ankeny, R. A. (2022). Philosophy of medicine. In. The Stanford Encyclopedia of Philosophy (Spring 2022 Edition), Edward N. Zalta (ed.), URL = <https://plato.stanford.edu/archives/spr2022/entries/medicine/>;.

Schramme, T. (2017). Philosophy of medicine and bioethics. In Schramme, T., & Edwards, S. (Eds.). Handbook of the Philosophy of Medicine. Springer Netherlands, pp. 3-15.

Stegenga, J. (2018). Care & Cure: An Introduction to Philosophy of Medicine. University of Chicago Press.

Thompson, R. P., & Upshur, R. E. (2018). Philosophy of medicine: An introduction. Routledge.

Gianina Druta[1,2]

[1]Oslo Metropolitan University, Norway; [2]Lucian Blaga University of Sibiu, Romania

The German tradition of staging Henrik Ibsen’s "Ghosts" and the tragic form

Wednesday, 05/Mar/2025 4:00pm - 4:20pm
ID: 203 / Session SP 02: 1
Short paper (abstract) | 15-minute presentation with a 5-minute Q&A
Keywords: Ibsen, Ghosts, German theatre, tragedy

The aim of this research is to investigate the main features of the German-speaking tradition of staging Henrik Ibsen’s *Ghosts* with a focus on the tragic approach to the play. Germany is the country that most performed *Ghosts* worldwide, whereas German is the language that has been most used to perform the play. This is the premise to turn to the German aesthetics to investigate its long-lasting tradition in staging *Ghosts* and its impact on the other theatre cultures’ approach of Ibsen and of this play specifically.

Methodologically and theoretically, the analysis combines the use of Digital Humanities tools such as graphs, maps and networks based on the Ibsen Stage database with theatre historiography tools. It also addresses the relationship between Ibsen and the concept of tragedy on the German-speaking stage.

Graph, map and network visualisations based on IbsenStage indicate a rich tradition of staging *Ghosts* as the most popular Ibsen play on the German-speaking stage.

Firstly, the graphs indicate three powerful moments in the German reception of *Ghosts*: the early reception (1894-1928); the post-war reception (1946-1950, 1969-1989); and the contemporary reception (1997-2016). Graphs also indicate the most influential institutions that contributed to the dissemination of the play during the abovementioned periods. Further contributor statistics based on IbsenStage indicate that the engagement with *Ghosts* of theatre practitioners like such as actors, directors and designers revolve around these institutions.

For instance, the early reception period is dominated by Deutsches Theater, Städtische Theater Leipzig and Das Ibsen-theater aus Berlin. These institutions/companies staged *Ghosts* mainly between 1899 and 1928 and reveal different approaches to Ibsen’s play ranging from naturalism/realism to expressionism. This variety of perspectives suggests also a different a treatment of the tragic aspect of the play, where the naturalist/realis vs. the expressionist lens come with different proposals to the audience.

The post-war period between 1946-1950 highlights the interest of the company of Albert and Else Bassermann in *Ghosts*, which they toured with in 50 different cities. The company takes one step further the naturalist and the expressionist legacy of staging *Ghosts* on the German stage as Albert Bassermann started his career under the guidance of Otto Brahm, but continued his training and acting career under Reinhardt’s guidance at Deutsches Theater, while also working at Lessing Theater. Else was his wife and her career was mostly tied to Max Reinhardt and Deutsches Theater.

The German-speaking tradition of staging *Ghosts* has a Swiss twist to it in 1980 with the company Bühne 64 taking the main stage. The company’s production reveals itself at the most powerful moment in the performance history of staging *Ghosts* on the German-speaking stage and an influential moment in the global reception of the play too in ‘80s.

The next powerful moment of *Ghosts* on the German-speaking stage takes place between 1997-2004 with the following two productions: A co-production of Theater des Ostens and Konzertdirektion Schlote (1997-2000) which marks the second most powerful moment in the reception of *Ghosts* on the German stage; and the production of Fränkisches Theater Schloss Maßbach (2004).

The interest in *Ghosts* decreases slightly on the German stage after 2006 which marks the Ibsen centenary and a moment of expansion in terms of productions worldwide. Yet two other companies make their contribution to disseminating *Ghosts*: Theater 58 (2009-2011) which marks another Swiss moment in the German-speaking reception of *Ghosts,* reaching mainly Swiss cities; and the company Markus&Markus (2015-2016) which points at recent trends in the approach of Ibsen on the German-speaking stage. Their production is part of a larger project entitled the Ibsen Trilogy which also includes *John Gabriel Borkman* and *Peer Gynt,* with *Ghosts* as the play that the company staged the most.

Secondly, the map visualisations indicate a decentralised geographical distribution, with the play being staged in both German, Austrian and Swiss cities. This also points at the impact of the touring activity of the private theatres and companies that disseminated the play across the German-speaking stage.

Thirdly, the networks highlight the complexity of the German tradition in which tensions between contrasting perspectives, as well as aesthetic intertwinings marked the staging of *Ghosts*, mainly until the end of WWII. The networks also point at the impact of the German approach to staging Ibsen on the European and American theatre traditions.

Finally, the visualisations highlight a complex mechanism of aesthetic transmission in the reception of *Ghosts* on the German stage with naturalism and expressionism as the two main tendencies as early as the end of the 20th century. Echoes of the tensions between naturalism and expressionism emerge also in the contemporary reception of the play in the 21st century. In analysing the expressionist legacy, the understanding of tragedy and its embodiment on stage is key for the analysis of the evolution of the tragic form not only in the German, but also more widely in the European and American reception of Ibsen.

Jens Edlund
KTH Royal Institute of Technology, Sweden

Workshop proposal: An exploration of speech-oriented research in non-speech-centric disciplines

ID: 272
Half-day conference-themed workshop
Keywords: speech sciene, speech technology

This workshop aims to explore speech-oriented research in non-speech-centric disciplines.

Research that involves speech and spoken interaction routinely encounters significant challenges and pitfalls in analysis and experiment design. Despite these hurdles, many fields rely heavily on understanding spoken communication, whether in oral history, social sciences, medical diagnostics, or the arts. In the areas of speech technology and speech science, at least some of these hurdles are known, as are some of their solutions. However, there remains a lack of collaboration and knowledge exchange between speech-centric disciplines and disciplines in which speech and spoken interaction is a central part, but not the topic of study per se.

The workshop will focus on two main strands: studies involving live participants, such as recordings or analysis of live interactions, and studies involving the analysis of existing data. The former involves considerable methodological as well as technical and ethical issues, while the main pitfalls in the latter involves the appropriateness and reliability of analysis methods and the reliability and appropriateness of the data given the research question.

The half-day workshop on "Speech-oriented research" recognises the need for greater coordination of interdisciplinary collaboration involving the analysis of or experimentation with speech and spoken interaction. Situated within the context of digital humanities, where traditionally non-technical fields increasingly engage with data and technology, this workshop brings together researchers from various domains to discuss the complexities of studies involving spoken human communication.

A preliminary half-day (afternoon) agenda includes the following, with roughly 1h per item, including breaks:

- An introduction to the workshop and its participants. Includes some examples of innovative collaborations (organisers)
- A keynote (invited external speaker)
- Poster presentations focussing on practical matters: experienced problems, successes, research questions, data sets. Submission and solicitation. Editorial review; choice based on a balanced set of presentations. (Accepted contributions.)
- An open discussion of themes gathered from the presentations. Structured questions for which we propose answers. (All, potentially in groups.)
- Interest in forum; interest in participation in white paper authoring, interest in participation in general organisation, interest in Dagstuhl proposal writing (all; the latter may be excluded from open discussion depending on number of participants).
- Concluding words. (short)
- Social event

1. Information about the organisers

Jens Edlund holds a PhD in Speech Communication and a Docent in Speech Technology. He is a full professor at KTH Royal Institute of Technology in Stockholm, and the Director of Språkbanken Tal, the speech branch of the Swedish National Research Infrastructure Nationella språkbanken, as well as the head of the KTH representation in the National Research Infrastructure HumInfra. He is responsible for CLARIN SPEECH, a CLARIN ERIC K-centre focusing on speech technology in the humanities and social sciences, and for the KTH participant in the Swedish Dariah membership.

Edlund's research is highly interdisciplinary, and he is currently the PI of a handful of interdisciplinary research projects ranging from accessibility research to conceptual development in parliamentary discourse. He has published well over 100 peer reviewed journal and conference articles with speech and spoken interaction is a common denominator in a range of disciplines.

Ambika Kirkland is a PhD student at KTH. One main interest is analysis of human perception of different types of speech, where she experiments with different technologies ranging from web-based perception experiments to EEG.

Axel Ekström is a post-doc In the University of Zurich. His work investigates the origin of speech from an acoustic and articulatory perspective, and spans humans, primates, and other mammals, and involves among other things attempts at clarifying terminology and methodology issues that arise when borrowing methods from one field to another.

Christina Tånnander is a speech technologist at the Swedish Agency for Accessible Media and an industrial PhD student at KTH. Her work includes evaluation methods that puts the human needs in the centre, and sits in the cross-section of several relatively disparate fields.

Ghazaleh Esfandiari is a PhD student at KTH. She studies multimodal, multiparty human interaction in for example meetings, borrowing questions and methods from several disciplines.

Jim O'Regan is a PhD student at KTH. He works on new methods of processing large amounts of existing speech, mainly from historical archives and/or under-resourced languages. He is involved in several projects where non-speech-centric researchers investigate speech-oriented phenomena.

2. Expected outcomes

The long-term goal is an increased awareness and willingness to collaborate on speech-oriented questions.

The practical goals are:

- A white paper. Workshop participants as well as external authors will be invited to participate.

- An international interest group (potentially, at some point, a formal SIG) with a structured forum (e.g. a mailing list).
- A Dagstuhl proposal to continue work on best practices and methods of information sharing between disciplines that avoids constraining individual disciplines.
- Finally, as a result of the Dagstuhl, a continued regular workshop.

Sara E. Ellis-Nilsson[1], **Anders Fröjmark**[1], **Terese Zachrisson**[2]
[1]Linnaeus University, Sweden; [2]University of Gothenburg, Sweden

From Parchment to Pixels: Building a Digital Research and Educational Resource for the study of Medieval Lived Religion in Sweden and Finland

Wednesday, 05/Mar/2025 4:00pm - 4:20pm
ID: 128 / Session SP 04: 1
Short paper (abstract) | 15-minute presentation with a 5-minute Q&A
Keywords: digitization, digital cultural heritage, digital research infrastructure, medieval studies, structuring historical data

This paper introduces the research and educational resource *Mapping Saints*, built by the research and digitization of cultural heritage project, *Mapping Lived Religion. The Medieval Cults of Saints in Sweden and Finland* (MLR).[1] The main of the project – and its main publication – was the creation of this digital resource to enable new, interdisciplinary approaches to the study the cults of saints. Moreover, the project was responsible for enabling the digitization of two cultural heritage collections: the analogue card catalogue, *Iconographical Index of Ecclesiastical Art in Sweden* (*Ikonografiska registret*) at the Swedish National Heritage board, and the collection of photographs of ecclesiastical art taken by Lennart Karlsson, held by the Swedish History Museum, and previously available as low-resolution images in the database *Medeltidens bildvärld* (Liepe and Ellis Nilsson 2021). The public interface was launched in November 2024.

The digital portal *Mapping Saints* contains a backend with a relational database and REST API, as well as a frontend with GIS map visualization. It was built in part by applying and testing the possibilities of linked (open) data. The paper will critically reflect on the digital methods applied in the project which are of importance to medieval religious studies, a long-established field. The project aimed to apply linked (open) data principles wherever possible to ensure sustainability. Thus, the resource uses unique internal identifiers and identifiers from authority databases, and it applies standardized vocabularies and structured data.

Applying research-driven digitization, the MLR-project worked with an overarching theoretical focus of "lived religion", that is how religion was *lived* or *done* (Ellis Nilsson et al. 2022). Rather than focussing on religious beliefs, this theoretical approach explores the practice of religion as it was manifested in the landscape and through the lens of the cults of saints. Of primary interest in this case are the expressions of the veneration of saints from 1164-1593, i.e. from the establishment of the medieval Uppsala church province, until its definitive end at the Uppsala Synod. Thus, it is this timeframe and research focus that users of the resource gain access to. In addition, the resource can even give insights into the period before and after the main period of investigation (the 17th and early 18th centuries), even though these results are fewer.

The foundation of the resource itself rests on its place register and the project's approaches to spatial analysis of past phenomena. The geographical boundaries of the project's research framework – the ecclesiastical province of Uppsala – provide the basis for the map design. This region included most of present-day Sweden, as well as Finland and the Karelian peninsula. This demarcation was made to better mirror medieval circumstances in analyses and geographical visualizations of the cult of saints in this period. Moreover, uniquely, even geographical points outside of this region are included if they are connected to lived religion undertaken by individuals from Sweden and Finland, for example, pilgrimages.

The methods involved in creating digital representations of medieval space included linking and combining diverse spatial data from digital and analogue resources. With some places, exact coordinates are available. GIS requires specific coordinates for all places; however, some places have uncertain coordinates, while some data is of a less exact and broader geographical area, such as a diocese. The map does not contain polygons to represent these latter types of areas; instead, it uses points, for instance, a diocese is a point next to the cathedral.

For places in Sweden, the initial development of the place register was partially achieved by harvesting and merging data from two databases: the Swedish National Heritage Board's database of archaeological sites and monuments *Fornsök* and *Bebyggelseregistret* (a register of built cultural heritage in Sweden). These provided the location of ruins and excavation sites and historical buildings, such as churches, as well as coordinates for archaeological finds of disused chapels and holy wells.

For Finnish places, the Finnish Heritage Agency's database (kyppi.fi, archaeological sites and building heritage), *Valtakunnallisesti merkittävien rakennettujen kulttuuriympäristöjen* (RKY), and Karl G. Leinberg's *Finlands territoriala församlingars namn, ålder, utbildning och utgrening* were first consulted to make it possible to compile places and their coordinates.

In a few cases, coordinates have also been determined based on the provenance of artefacts (e.g. objects and parchment fragments) and spatial references from medieval narratives (e.g. miracle stories).

As part of the development process, it was found necessary to create new analytical concepts, such as 'cult manifestation'. This refers to when evidence for a saint's cult is manifest, represented by a particular item, in a particular location, and during a specific time-period. It can refer to physical objects, buildings, and landscape features, as well as immaterial phenomena such as narratives, feast days, and devotional acts (Ellis Nilsson et al. 2023). As cult manifestations and places are central tenets to the project's research, the database is structured around these two database tables.

Moreover, in order to deal with the plethora of dates and dating practices for the various source categories, it was necessary to develop a concept and common denominator. Termed the *functional period*, this field indicates the date – usually an interval – when the cult manifestation was active. The min-max from this field is connected to the resource's time-slider, enabling temporal visualizations and analysis.

The paper concludes by providing two concrete cases which show the usefulness of *Mapping Saints* for research into lived religion: the inclusion of altars as unique places and the analytical possibilities of identifying and mapping previously unknown medieval individuals. In the future, this resource can be linked with other digital resources – in particular, 'linkable silos' – created by research-driven projects. Throughout, the paper will also critically assess the place of *Mapping Saints* in the wider context of the creation of bespoke tools for research projects in the digital humanities.

[1] Acknowledgement is hereby made to the contributions of the core MLR project members listed here in addition to those participating at the conference: Lena Liepe, Sofia Lahti, Vilma Mättö, Steffen Hope, Benjamin Allport, Johan Åhlfeldt, Julia Beck, and Kristin Åkerlund.

Bibliography

Ellis Nilsson, S., Zachrisson, T., Fröjmark, A., Liepe, L. and Åhlfeldt, J. (2023). “Mapping Saints: creating a digital spatial research infrastructure to study medieval lived religion”, in Alexandra Petrulevich and Simon Skovgaard Boeck (eds.). Digital Spatial Infrastructures and Worldviews in Pre-Modern Societies. York: ARC Humanities Press, 33-58.

Ellis Nilsson, S., Liepe, L. and Zachrisson, T. (2022). “Scandia introducerar: Levd religion i det förmoderna Nordeuropa” (Scandia introduces: Lived Religion in Pre-modern Northern Europe). Scandia 88:2: 317–337.

Liepe, L. and Ellis Nilsson, S. (2021). “Medieval Iconography in the Digital Age: Creating a Database of the Cults of Saints in Medieval Sweden and Finland”, Iconographisk Post: Nordisk tidskrift för bildtolkning/Nordic Review of Iconography, 2: 45-63.

Dāvis Eņģelis[1,3], Valdis Saulespurēns[2], Haralds Matulis[1,4], Anda Baklāne[2]

[1]Institute of Literature, Folklore and Art of the University of Latvia (ILFA), Latvia; [2]Latvian National Library (LNB); [3]Jāzeps Vītols Latvian Academy of Music; [4]University of Latvia, Faculty of Humanities

Conceptual Challenges in Creating a Corpus of Latvian Music Texts from Digitized Newspaper Archive

Thursday, 06/Mar/2025 2:00pm - 2:20pm
ID: 243 / Session SP 05: 1
Short paper (full-text) | 15-minute presentation with a 5-minute Q&A
Keywords: digital archives, music periodicals, musicology, digital collections

To create the first digitally accessible corpus of Latvian music texts, we used text aggregation from an archive of digitized Latvian newspapers at the Latvian National Library (LNB). By broadening the definition of what counts as a music text, we aimed to facilitate access to texts about music located outside of the known musical periodical editions and genres of music journalism.

We address several research problems: facilitation of access to digitized sources of Latvian musical periodicals; integration of digital humanities into the discourse of Latvian historical musicology; promotion of explicit description of the process of data acquisition from digital sources.

The timescale of the corpus covers accounts of music texts published in the 19th century up to musical thought from modern newspapers. To select the primary collection of texts for candidates to include in the corpus, we used a music related keyword (f=147) list from manually compiled pilot corpora. By close reading of findings we finetuned the search algorithm in order to reduce the amount of false positive results.

In our paper, we address the conceptual issue of defining what is a text about music, alongside more customary obstacles of corpus curation – inconsistent orthography, and data use restrictions. Discussing preliminary results, we touch upon the further potential directions of research, among them, the inclusion of generative large language model API in the workflow, which was experimented in the research to automate the typology of music text genres, but was beyond the primary goal of the current research.

Valts Ernštreits
University of Latvia Livonian Institute, Latvia

Transforming Livonian Folklore Archives into a Multifaceted Digital Resource

ID: 301 / **WS10: 5**
Tradition Archives Meet Digital Humanities II
Keywords: Livonian, Folklore, Corpus, Community Involvement

Endangered languages and cultures are often characterized by fragmented or insufficient documentation, scattered archives, and limited resources—both financial and human—for processing and digitizing collections to transform them into digital resources and tools. Such communities are also profoundly influenced by majority cultures, often resulting in a decline in the number of proficient speakers and limited access to archival materials, such as folklore, that are "locked" behind language barriers.

Livonian, an indigenous language of Latvia, is no exception. With a community of approximately 1,000 members scattered across Latvia and fewer than 20 fluent speakers, Livonian represents a critically endangered language facing significant challenges. While there is a growing interest in language acquisition and heritage reclamation, access to these resources remains a key challenge that digital tools and resources could address.

One of the most significant intangible heritage collections for Livonian is the handwritten collection of Livonian legends and folk tales housed at the Estonian Literary Museum. Collected during the 1920s and 1930s, this material originates from a time when Livonian was still an everyday language within a cohesive community. This collection, which accounts for nearly one-third of all written Livonian texts, is not only invaluable as a repository of intangible heritage but also as a linguistic resource, bridging classical written texts and spoken language.

Over recent decades, significant efforts have been undertaken at the University of Latvia Livonian Institute to digitize and integrate this collection into a digital Livonian language corpus. Central to this initiative are approaches that ensure the multifaceted use of the created resources. These include the development of a corpus, vocabulary and morphology data, text collections, and resources for building machine translation and speech solutions, as well as tools that enable non-proficient users to access the collections' contents and overcome language barrier.

Crucially, the involvement of fluent Livonian speakers has been integral to this work. Their expertise has been applied to the transliteration of texts, the provision of sound recordings, annotations, and raw translations, making the collection more comprehensible, audible, and usable not only for researchers but also for language learners and those seeking to build a Livonian language environment. At the same time, it has provided speakers with the rare opportunity to immerse themselves in a Livonian language and heritage environment—an invaluable experience given the small and scattered number of speakers. This initiative allows them to practice the language actively and contribute to the community.

This presentation will describe approaches to creating multi-purpose and resource-efficient digital resources and tools that serve both the academic community and the needs of an endangered language community.

Ghazaleh Esfandiari Baiat, **Edlund Jens**
Royal Institute of Technology (KTH), Sweden

Focusing on the task at hand: Bespoke annotation and modelling of ranking tasks in meetings

Wednesday, 05/Mar/2025 5:20pm - 5:40pm
ID: 277 / Session SP 02: 5
Short paper (full-text) | 15-minute presentation with a 5-minute Q&A
Keywords: MEET corpus, three-party meetings, ranking tasks, modelling interaction

The MEET corpus is a collection of annotated recordings of three-party meetings in which the participants took part in collaborative decision-making. We present work where social science meets speech science and speech technology and demonstrate how a bespoke annotation scheme combined with programmatic and task-dependent interpretations of the concepts “proposition”, “question”, and “decision” affords an incremental and operationalised view of the process of arriving at group consensus. The outcome of this process is an incremental and dynamic model that, although it is designed to target a specific set of research questions, can serve as a basis for a wide range of investigations. The model is able to show a snapshot of the current state and the propositions and decision points that led there at any given time in the conversation, and can be extended to a range of other, similar tasks.

Pascale Feldkamp
Center for Humanities Computing, Aarhus University, Denmark

Dynamics of Literary Fields: Tracing Influence of Danish Canonical Novels (1870-1900)

ID: 309 / WS04A: 1
Explorations of the dynamics of cultural phenomena in text corpora
Keywords: literary studies, canon, literary influence, document embeddings, stylistics

Recently, work in computational literary studies has examined cultural change through the similarity of literary works across time (Barr´e 2024; Griebel et al. 2024; Liddle 2019). Along the same lines, this presentation examines how Danish literature evolved during the Modern Breakthrough period (1870–1899) through shifts in similarity, focusing on literary influence between works that are canonical today and the broader literary field of the time.

This work focuses on novels' textual profiles—specifically represented by text embeddings—to trace shifts in textual similarity across novels. These profiles allow us to connect the historical transition to Realism in the Modern Breakthrough with the internal dynamics of the literary field, offering a data-driven perspective on canonical and non-canonical evolution.

Previous work seeking to distinguish canonical and non-canonical works has predominantly relied on feature engineering, using linguistic metrics and information theory to show, for example, how canonical texts exhibit higher levels of reading difficulty than more popular works of literature (Algee-Hewitt et al. 2016; Bizzoni et al. 2024; Wu et al. 2024). However, few studies go beyond selected features of the stylistic dimension when examining the canon. To capture the multidimensional nature of literature, we use doc- ument embeddings, which encode stylistic and semantic features of a novel at various levels (Wang et al. 2023; Terreau et al. 2024; Reimers and Gurevych 2019).

This work draws on a corpus of 839 novels (including 114 canon novels) representing the production of Danish novels in the period. Using diachronic comparisons with rolling windows and cosine similarity, we show that internal diversity within both canonical and non-canonical categories increased over time. This reflects a more streamlined literary culture aligned with the Modern Breakthrough, as genres like the historical novel declined and Realism emerged as a dominant trend. Moreover, we show how canonical novels initially stood out with a distinct textual profile, which non-canonical novels gradually approximated, making the canonical novels appear as "trendsetters".

As embeddings are inherently opaque, we supplement our analysis with more trans-parent stylistic features – including word count, textual diversity, word length, sentence length, and readability measures. This parallel analysis corroborates the findings from our embeddings-based approach and suggests connections between abstract document representations and tangible stylistic characteristics.

Ultimately, this study demonstrates the utility of embeddings in tracing changes in literary culture, providing insights into the canon's role within the dynamics of a broader literary field.

Bibliography

Algee-Hewitt, Mark, Sarah Allison, Marissa Gemma, Ryan Heuser, Franco Moretti, and Hannah Walser (2016). Canon/Archive. Large-scale Dynamics in the Literary Field. Stanford Literary Lab.

Barré, Jean (2024). "Latent Structures of Intertextuality in French Fiction: How literary recognition and subgenres are framing textuality". In: Computational Humanities Research Conference. Aarhus, Denmark: CEUR Workshop Proceedings, pp. 21–36.

Bizzoni, Yuri, Pascale Feldkamp, Ida Marie Lassen, Mia Jacobsen, Mads Rosendahl Thomsen, and Kristoffer Nielbo (Apr. 2024). Good Books Are Complex Matters: Gauging Complexity Profiles Across Diverse Categories of Perceived Literary Quality. doi: 10.48550/arXiv.2404.04022. arXiv: 2404.04022 [cs].

Griebel, Sarah, Becca Cohen, Lucian Li, Jaihyun Park, Jiayu Liu, Jana Perkins, and Ted Underwood (2024). "Locating the Leading Edge of Cultural Change". In: Computational Humanities Research Conference. Aarhus, Denmark: CEUR Workshop Proceedings, pp. 232–245.

Liddle, Dallas (2019). "Could Fiction Have an Information History? Statistical Probability and the Rise of the Novel". en. In: Journal of Cultural Analytics, p. 22. doi: 10.22148/16.033.

Reimers, Nils and Iryna Gurevych (2019). "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks". In: Conference on Empirical Methods in Natural Language Processing. url: https://api.semanticscholar.org/CorpusID:201646309.
Terreau, Enzo, Antoine Gourru, and Julien Velcin (July 2024). Capturing Style in Author and Document Representation. en. arXiv:2407.13358 [cs]. url: http://arxiv.org/abs/2407.13358

Wang, Andrew, Cristina Aggazzotti, Rebecca Kotula, Rafael Rivera Soto, Marcus Bishop, and Nicholas Andrews (2023). "Can Authorship Representation Learning Capture Stylistic Features?" In: Transactions of the Association for Computational Linguistics 11. Place: Cambridge, MA Publisher: MIT Press, pp. 1416–1431. doi: 10 . 1162 /tacl_a_00610. url: https://aclanthology.org/2023.tacl- 1.80

Wu, Yara, Yuri Bizzoni, Pascale Moreira, and Kristoffer Nielbo (Mar. 2024). "Perplexing Canon: A study on GPT-based perplexity of canonical and non-canonical literary works". In: Proceedings of the 8th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature (LaTeCH-CLfL 2024). Ed. by Yuri Bizzoni, Stefania Degaetano-Ortlieb, Anna Kazantseva, and Stan Szpakowicz. St. Julians, Malta: Association for Computational Linguistics, pp. 172–184. url: https://aclanthology.org/2024.latechclfl-1.16.

Elena Fernández Fernández, Jana Meier, Simon Clematide
University of Zurich

The Speed of the West? Measuring Social Acceleration in World Capitals using Newspapers and Quantitative Methods (1999-2018)

ID: 290 / WS04A: 6
Explorations of the dynamics of cultural phenomena in text corpora
Topics: Communication studies, Computer science, Cultural studies, text, data mining and analysis, data modeling
Keywords: Text Data Mining, World City Theory, Science and Technology in Society, Media and Communication Studies

World cities are defined by many scholars not only as major urban settlements found around the globe, but as influence hubs of economic, cultural, geographic, and technological innovation. It has been proposed by many authors (Sassen, Castells, Knox) that they conform an independent ecosystem, stranged from their respective national territories cultural norms. In this article we explore the possibility of detecting a world city ecosystem under the paradigm of information behaviour, yet challening geographic location as the only determining factor. Thus we use two types of news outlets placed at eight world cities internationally as a proxy: 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. Our observational time covers twenty years (1999-2018). We frame our analysis of information behaviour under the Theory of Social Acceleration, that states that rush is a side effect of increasing waves of technology arriving to society. By defining narrative units as triplets of Subject-Verb-Object (in dialogue with Franzosi and Fernández Fernández et al.), we quantify the number of SVO triplets found in our dataset of newspapers across our observational time, seeking to detect whether it is possible to find similar rates in all our newspapers, as they are all headquartered in world cities (London, Dublin, Paris, Berlin, Zurich, Turin, Madrid, and Chicago). Our results show a shared information ecosystem formed by the seven newspapers of record where it is possible to detect incrementing rates of SVO triplets over time, and therefore, empirically validating the theory of social acceleration. However, Chicago Daily Herald, a suburban newspaper headquartered in Chicago, and therefore, also located at a world city, shows patterns of speed deceleration (decreasing numbers of SVO triplets over time), consequently challenging the geographic location as the only factor to consider in the formulation of the world city theory (at least from an information perspective).

Bibliography

Castells, Manuel. The Rise of the Network Society. Blackwell Publishers, 2000.

Fernández Fernández, Elena, et al. “Measuring the Acceleration of the Social Construction of Time using the BOE (Boletín Oficial del Estado)”. CHR 2020: Workshop on Computational Humanities Research, November 18-20, Amsterdam, The Netherlands.

Franzosi, Robert. Quantitative Narrative Analysis. Sage, 2010.

Knox, Paul L. “World Cities in a World-System”. World Cities in a World System, edited by Paul L. Knox and Peter J. Taylor, Cambridge University Press, 1995, pp. 3-20, 2014, doi: 10.1093/llc/fqu057.

Sassen, Saskia. The Global City. New York, London, Tokyo. Princeton University Press, 2001.

Elena Fernández Fernández, Jana Meier, Simon Clematide
University of Zurich

Hands-on tutorial on „The Speed of the West“ methodology

ID: 318 / WS04B: 1
Explorations of the dynamics of cultural phenomena in text corpora
Keywords: Text Data Mining, World City Theory, Science and Technology in Society, Media and Communication Studies

The tutorial introduces the methodology used for the presentation "The Speed of the West" by the same authors.

Mats Fridlund
University of Gothenburg, Sweden, Sweden

Towards Responsible DH: Practices of Critical and Caring DH

ID: 253 / **Panel 2: 1**
60-minute panel
Keywords: responsible DH, activism, interdisciplinarity, project collaboration, ethics

Introduction

In line with the theme of DHNB 2025, Digital Dreams and Practices, this panel will explore some of the key ethical issues inherent in and raised by DH practices. The panel strives to initiate a reflective and critical discussion on the challenges of practicing what we call 'Responsible DH' - DH practices in academia and the GLAM sector that strive to explicitly consider the ethical, social and political challenges of representation, sustainability, fair use of data, bias and inequalities. The panel will also examine the societal and environmental impacts of the use of emerging technologies, including artificial intelligence and large-scale data analysis.

With vast amounts of data being digitized, shared, and analyzed, questions of who owns this data, who has access to it, and how it should be used becomes increasingly central (Proferes, 2020). The digitization of cultural artifacts and texts also raises questions about who is represented in digital spaces and how and by whom. This has led to concerns about digital colonialism, where DH practices may reproduce historical imbalances of power, particularly in the representation of non-Western cultures (Stobiecka, 2020; Risam, 2018). Such questions regarding power imbalances become particularly pertinent in projects where Western researchers and developers collaborate in projects with partners in the Global South or that represent marginalized groups.

In addition, the increasing reliance on algorithms and machine learning in digital humanities projects raises questions about discrimination and bias. For example, text-mining projects may inadvertently favor certain types of language, literature, or data over others, thus reinforcing existing societal biases (Fiormonte, 2012). Last but not least, digital projects often require significant resources in terms of energy and water, raising concerns about their environmental impact and sustainability (Nowviskie, 2015).

The panel discussion aims to bring together academics and practitioners that through their positions and experience of participating in 'responsible DH' projects will provide different complementary perspectives on these issues.

Expanding on these points, the panel will delve deeper into the ethical implications of data ownership and access. As data becomes a valuable commodity, the question of who controls this data and the power dynamics involved in its use becomes crucial. (Allington et al. 2016)The panel will discuss how to ensure equitable access to data and prevent monopolization by a few entities.

Furthermore, the representation of diverse cultures in digital spaces will be scrutinized. The panel will explore strategies to avoid digital colonialism and ensure that digitization efforts do not perpetuate historical injustices. This includes discussing best practices for collaborating with partners in the Global South and marginalized communities to ensure their voices are heard and respected.

The role of algorithms and machine learning in DH projects will also be a focal point. The panel will examine how these technologies can inadvertently reinforce biases and what measures can be taken to mitigate such risks. This includes discussing the importance of transparency in algorithm design and the need for diverse datasets to train these systems.

Finally, the environmental impact of digital projects will be addressed. The panel discussion will touch upon ways to make DH practices more sustainable, such as using energy-efficient technologies and considering the lifecycle of digital projects. By bringing together a diverse group of experts, the panel aims to foster a comprehensive discussion on how to practice responsible DH in a way that is ethical, inclusive, and sustainable.

Panel structure (90 minutes in total)

1. **Introduction (10 minutes):** The moderator (Mats Fridlund) will set the stage, providing an overview of the panel's theme and introducing key discussion points around responsible DH.

1. **Individual presentations (5-7 minutes each):** Each panelist will provide a short presentation reflecting on their experiences and insights regarding ethical issues in DH within their areas of DH practice. These presentations will focus on the following questions:

- What ethical issues and potential corresponding strategies are involved in developing DH projects (e.g., collecting and training on data)?
- How can we address imbalances in collaborations between different institutions or regions?
- What are the environmental sustainability concerns, including energy and water consumption, of digital projects?
- What principles or approaches should guide the development of responsible DH practices?

Joint discussion (45 minutes):
A joint discussion where panelists will engage with key questions, discussing potential strategies, principles or interventions for addressing ethical challenges in DH practices.

Outlines of individual contributions

- **Dorna Behdadi** (University of Gothenburg): PhD Practical philosophy, Researcher in AI Ethics at the Gothenburg Research Infrastructure in Digital Humanities (GRIDH). This presentation will focus on providing a general outline of various ethical questions raised by, or in conjunction with, DH practices, as well as charting various strategies or approaches for how one may address such challenges.

- **Valerie Shafer** (University of Luxembourg): Professor in Contemporary European History at the C²DH (Luxembourg Centre for Contemporary and Digital History) and Associate Researcher at the Center for Internet and Society (CIS – CNRS UPR 2000). This presentation will explore how researchers can balance the objectives of historical documentation with ethical considerations in digital heritage research, particularly when dealing with personal data and born-digital materials.
- **Marianne Ping Huang** (Aarhus University): Associate Professor at School for Communication and Culture.
- **Jonathan Westin** (University of Gothenburg): Associate Professor in Conservation, Acting Director of Gothenburg Research Infrastructure in Digital Humanities, and leads the project "DIGICURE: Digital Cultural Resilience and Protection" in Ukraine.

Bibliography

Allington, Daniel, Sarah Brouillette, and David Golumbia. "Neoliberal tools (and archives): a political history of digital humanities." Los Angeles Review of Books (2016): 35-78.

Fiormonte, D. (2012). Towards a cultural critique of the digital humanities. Historical Social Research/Historische Sozialforschung, 59-76.

Nowviskie, B. (2015). Digital humanities in the Anthropocene. Digital Scholarship in the Humanities, 30(suppl_1), i4-i15.
Proferes, N. (2020). What ethics can offer the digital humanities and what the digital humanities can offer ethics. In Routledge International Handbook of Research Methods in Digital Humanities (pp. 416-427). Routledge.

Risam, R. (2018). Decolonizing the digital humanities in theory and practice. In The Routledge companion to media studies and digital humanities (pp. 78-86). Routledge.

Stobiecka, M. (2020). Archaeological heritage in the age of digital colonialism. Archaeological Dialogues, 27(2), 113-125.

Mats Fridlund, Daniel Brodén, Michael McGuire
University of Gothenburg, Sweden, Sweden

Triangulating Terrorism through Text Mining: A Comparative Study of Terrorism Discourses in Swedish Cultural Periodicals during the Cold War

Thursday, 06/Mar/2025 9:30am - 10:00am
ID: 264 / Session LP 04: 2
Long paper (abstract) | 20-minute presentation with a 10-minute Q&A
Keywords: terrorism, intellectual history, text mining, corpus studies, word rain

1. Introduction

The threat of 'international terrorism' became an urgent issue in Swedish and Western public discourse around 1970 (Hansén 2007; Brodén et al. 2023). At the time, the violent tactics used by sub-state militants drew significant international attention. In fact, research has shown that it was not until then that the modern notion of terrorism emerged, drawing together a range of violent tactics (aircraft hijackings, political assassinations, bombings, hostage-taking, etc.) in another way than before (Stampnitzky 2013; Zoller 2018). Rather than as brutal political violence in local conflicts, primarily in Global South, terrorism became interpreted as a specific transnational threat, epitomised by the Palestinian skyjackings starting in 1968 and the attack against Israeli Olympic team members at the 1972 Summer Games in Munich. Now, terrorism was perceived as impacting the foundations of modern societal order, affecting the security of authorities and civilians as well as diplomatic relations and global communication networks. However, different public forums, such as newspapers, political debates, etc., continued to interpret the concept of terrorism in diverse ways, producing different discourses that reflected their particular cultural frames, agendas and audiences.

2. Purpose and aims

The project connects to the conference call for proposals to explore "the integration of traditional humanities scholarship with computational techniques" through an intellectual history (history of ideas) and history of concepts study that uses digital text mining and methods from corpus linguistics to conduct a comparative historical study of the political concepts of 'terror' and 'terrorism' during the Cold War. In particular it marries the traditional approach of studying the development of a periode's (t*he Cold War*) cultural discourse is shaped through analysing the public debate regarding a cultural issue or concept (*terrorism*) on a significant national public area (*Swedish cultural periodical*s) where involves prominent authors, public intellectuals, and politicians, with the use computational methods. This takes its inspiration from and contributes to recent development of the use of digital methods within intellectual history (see among others: Edelstein 2016, Hill 2016, London 2016, Tolonen et al. 2021, de Bolla 2023).

Building on a previous pilot study (Fridlund et al. 2023) presented at DHNB2023, this paper employs text mining to investigate the discourse on terrorism in Swedish periodicals during the Cold War, 1945–1991, providing a comprehensive perspective on the emergence of the modern notion of terrorism. While Sweden maintained a national self-image as a peaceful contrast to the otherwise conflict-ridden world of the Post War years (Salomon et al. 2004; Cronqvist 2015), 'international terrorism' emerged as a domestic threat in the 1970s, manifested by politically motivated violence carried out by Croatian separatists and militants linked to the Red Army Faction (RAF). Nevertheless, the term 'terrorism' was presented in the broader public discourse throughout the entire Cold War era (Hansén 2007; Brodén et al. 2023).

The proposed paper aims to triangulate and compare the discourse on terrorism during the Cold War across three Swedish periodicals, each representing distinct different ideological perspectives: the conservative *Svensk Tidskrift, the Zionist Judisk tidskrift ('Jewish Periodical')* and *Bonniers Litterära Magasin, BLM ('Bonnier's Literary Magazine')*, a prominent national cultural periodical that took a leftward turn in the mid-1960s. By examining there diverse periodicals, the paper will contribute to 1) a more nuanced understanding of the terrorism discourse in Sweden, demonstrating the analytical value of comparing discourse across different periodicals, and; 2) the methodological advancement of journalism-oriented text mining projects in the Digital Humanities, which have predominantly focused on newspapers rather than periodicals.

The paper is driven by two general straightforward research questions: What specific discourses of terrorism appear in the three periodicals? In what ways do these different discourses overlap?

The paper will demonstrate how a mixed-methods approach, combining Corpus Linguistics Digital History and Media History, can deepen our understanding of the transformation of the concept of terrorism during the Cold War period. On a higher level, the paper feeds into the debates within Digital Humanities about the importance of pursuing intersections between interpretative and computational analysis through a collaborative process in which humanities scholars and data analysts engage in a dialectical relationship (Rockwell & Sinclair 2016).

3. Material

This study's source material has two distinctive features compared to most other digital projects that focus on newsprint. First, the project examines periodicals (specifically 'cultural periodicals') rather than newspapers. Second, instead of relying on publicly available online corpora of digitised newsprint, we created, annotated, and searched our own custom-made corpora, consisting of selected relevant articles.

Regarding the details of the three corpora. The corpus consists of all articles from the three journals from the period 1945-1991 containing one or more occurrences of "terrorism", "terrorist", and/or "terror." Each journal corpus also contains article metadata and can be searched as sub-corpora, separating terrorism/terrorist and terror related articles as well as specific time-periods during the Cold War. The three cultural periodicals are: *Svensk Tidskrift:* with 189 terror-relevant articles (75 containing terrorism nd 114 containing terror) of in total 476,338 words; *Judisk Krönika:* with 273 articles (116 on terrorism and 157 on terror) of 330,909 words: and BLM - *Bonniers Litterära Magasin*: 209 articles (74 on terrorism and 145 on terror) of 619,873 words.

4. Method

Comparative research on discourses in different newspapers and periodicals is an established field within the humanities and social sciences (van Dijk 1991; Machin & Mayr 2012). While such studies typically use discourse analysis and content analysis to examine how various media outlets frame a particular topic, tracing differences in language and perspective, computationally-driven approaches have been rare, despite their potential to quantify diverging and converging patterns across multiple sources at scale. In the field of terrorism studies, some quantitative approaches to comparative discourse analysis have been undertaken (Ditrych 2014), but these efforts have also stayed within traditional frameworks.

Utilising both distant and close reading, we will examine data visualisations of the different terrorism discourse in *Svensk Tidskrift, Judisk tidskrift* and *Bonniers Litterära Magasin* in the form of:

- Heat maps for Named Entities – people and countries
- Tables – Named Entities – people and organisations
- Tables – Word Frequencies of terrorism-related words
- Collocations of words frequently occurring with terrorism
- Change and distribution of word frequency across time (by decade and before/after 1968)
- Comparison of terrorism/terror sub-corpora within each journal and across journals
- Word rain (visualisations combining word frequency and semantic clustering)

While these approaches can be considered traditional, we will also incorporate the analysis of word frequencies based on the new ‘word rain’ approach (Skeppstedt et al. 2024) that overcomes the lack of a semantically motivated positioning of the words in conventional word clouds..

5. Analysis

In this paper, we examine the usage of the terms “terrorism,” “terrorist,” and “terror” across different journals and over time. We also investigate the evolution of terrorism as a modern concept. Our earlier pilot study revealed that in the 1940s and 1950s, “terrorism” most commonly referred to state actors such as Nazi Germany or the Stalinist Soviet Union and was often used similarly to “terror.” However, in the 1960s and 1970s, there was a notable shift where “terrorism” more frequently referred to non-state individuals and groups. This shift is evident in all three journals but occurs earlier in Judisk Krönika, becoming prominent in the 1960s, whereas in *BLM* and *Svensk Tidskrift*, the shift is more noticeable in the 1970s. *Judisk Krönika* is also unique in that even in the late 1940s, “terrorist” is used to refer to individuals and non-state groups.

We also compare mentions of locations associated with terrorism and terror-related activities, revealing interesting, though expected, differences across the three journals. All journals show strong concentrations around Europe. *Svensk Tidskrift* has the broadest worldwide distribution, reflecting references to decolonization-related terrorism in Africa, the Americas, and Asia. *Judisk Krönika* has a strong focus on Israel and the Middle East. *BLM* has a wide distribution with concentrations in Africa but also mentions many places as abstract concepts such as “Hollywood” and “Broadway.”

Furthermore, we compare differences in terrorism discourse across journals in relation to Cold War concepts of communism and socialism. These concepts are significant in both *Svensk Tidskrift* and *BLM*, but they are discussed very differently in the context of terrorism. In *Svensk Tidskrift,* communism and socialism are often conflated and almost always directly connected to terrorism and terror in a negative context, referring to both state and non-state actors. In BLM, communism and socialism are more frequently discussed as general social, cultural, and literary concepts, not always negatively as in *Svensk Tidskrift.* By comparison, communism and socialism are mentioned less directly in the context of terrorism in *Judisk Krönika*. Instead, terrorism discourse in *Judisk Krönika* is most strongly focused on Palestinian and Arab-related terrorism against Israel, with “arabisk terrorist” being the most frequent collocation.

6. Conclusions

In our paper, we explore the cultural concept of terrorism within a Digital Humanities (DH) context from a novel perspective, employing a mixed methods approach. We combine comparative corpus-based statistical text analysis with visualizations of a unique class of print corpora to investigate how people, places, and concepts are connected to terrorism. This involves utilizing both traditional methods from corpus linguistics and data visualization, as well as the innovative data visualization and analysis method known as ‘word rain’.

Our analysis and preliminary results indicate that the journals selected for this study each represent different facets of the cultural imaginary of terrorism, enabling us to capture a comprehensive picture of terrorism within social, cultural, and political contexts. Despite noticeable and expected differences in content and discourse, these journals also exhibit similarities and common themes. All journals demonstrate a shift in the concept of terrorism to predominantly refer to non-state actors during the late 1960s and 1970s. This shift is mirrored in popular culture, where references to terrorism increased in literature, film, and art, reflecting changes in cultural reality, perception, and representation of terrorism.

The journal *BLM* explores terrorism from a cultural perspective, with frequent references to literature, art, and philosophy. It also employs historical references to connect modern concepts of terrorism with the past. *Svensk Tidskrift* offers a more conservative political perspective, with discourse reflecting anti-communist and anti-socialist views. *Judisk Krönika* provides a Swedish Jewish perspective.

Taken together, we believe that a comparative and contrastive study of these three journals will offer valuable insights into the place of terrorism in Swedish cultural debates during the Cold War. Additionally, this study demonstrates the utility of computational methods in the fields of intellectual history and media history.

Bibliography

Brodén, Daniel, Mats Fridlund, Leif-Jöran Olsson, Magnus P. Ängsal, Patrik Öhberg (2023): ‘The Diachrony of the New Political Terrorism: Tracing Neologisms and Frequencies of Terror-related Terms in Swedish Parliamentary Data 1971–2018’, DHNB 2022: Proceedings, CEUR-WS.

De Bolla, Peter, ed. Explorations in the Digital History of Ideas: New Methods and Computational Approaches. Cambridge University Press, 2023.

Ditrych, O (2014): Tracing the discourses of terrorism, Palgrave Macmillan.

Edelstein, Dan. "Intellectual history and digital humanities." Modern Intellectual History 13.1 (2016): 237-246.

Fridlund, Mats (2023) Michael Azar, Daniel Brodén, & Michael McGuire, “The Cultural Imaginary of Terrorism: Close and Distant Readings of Political Terror in Swedish News and Fiction During the Cold War”, Digital Humanities in the Nordic and Baltic Countries Publications 5 (2023):1, 90–104

Hansén, D (2007): Crisis and perspectives on policy change. Stockholm: Försvarshögskolan.
Hill, Mark J. "Invisible interpretations: reflections on the digital humanities and intellectual history." Global Intellectual History 1.2 (2016): 130-150

Jennifer London, “Re-imagining the Cambridge School in the Age of Digital Humanities”, Annual Review of Political Science 19, 1 (2016): 351-373.

Machin, D & A Mayr (2012): How to do critical discourse analysis: A multimodal introduction, London: SAGE.
Rockwell, G, & Sinclair, S (2016): ‘Thinking-through the history of computer-assisted text analysis'. In C Crompton, et al., eds., Doing digital humanities. London: Routledge.

Skeppstedt, M, M Ahltorp, K Kucher, Mats Lindström (2024): ‘From word clouds to Word Rain: Revisiting the classic word cloud to visualize climate change texts’, Information Visualization; 1–22.

Stampnitzky, L (2013): Discipling terror, Cambridge: Cambridge University Press.

Tolonen, Mikko, Mark J. Hill, Ali Zeeshan Ijaz, Ville Vaara, and Leo Lahti. "Examining the early modern canon: The English Short Title Catalogue and large-scale patterns of cultural production." Data visualization in enlightenment literature and culture (2021): 63-119.

van Dijk, T A (1991): Racism and the press, London: Routledge.

Zoller, Silke (2021): To deter and punish: Global collaboration against terrorism in the 1970s, Columbia University press.

Stefan Gelfgren[1], Jakob Dahlbacka[2], Bo Ejstrud[3], Andreas Tjomsland[4,5]
[1]Umeå University, Sweden; [2]Åbo Akademi University, Finland; [3]Holstebro Museum, Denmark; [4]Volda University College, Norway; [5]University of Agder, Norway

Nordic Places of Worship (NordPoW) – an example of the need for solid humanist competencies when building research infrastructures

Friday, 07/Mar/2025 1:30pm - 2:00pm
ID: 185 / Session LP/SP 10: 1
Long paper (full-text) | 20-minute presentation with a 10-minute Q&A
Keywords: church history, religious geography, sociology of religion, data visualization, GIS

Using the so-called NordPoW map (a joint Nordic map over churches and prayer houses) as the starting point, this paper discusses what is mentioned in the call for papers, namely that: “Digital humanities begin with explorations of data with a humanities lens, but the strongest impact is achieved when solid computational methods are applied to questions important to established fields with long research traditions.” It will be argued that there is no hierarchy between the “humanities lens” and “computational methods” – both are equally important for a successful project, and mutual communication, intertwined interests, and a shared understanding of the project are key. What is crucial, and this cannot be stressed enough, for a research-oriented digital humanities project, is the research question.

We have built a scalable infrastructure for research into the religious geography of the Nordic countries, today with 6000 objects (but we expect it to grow to at least 20.000 objects).

The research questions and the discussions on how to curate the data are grounded in our different disciplines and competencies, rooted in traditional humanities. The NordPoW project group consists of four researchers with backgrounds in the sociology of religion, history, archaeology, theology, church history, and digital humanities. We have extended and close collaboration with system developers with GIS expertise from Humlab, Umeå University, and the Swedish research infrastructure InfraVis. This project would not have been possible without a strong background and competencies in traditional humanities disciplines. The paper will discuss the collaboration and experiences from the project.

Koraljka Golub
Linnaeus University, Sweden

Challenges in AI: Indexing LGBTQ+ fiction

Wednesday, 05/Mar/2025 5:00pm - 5:20pm
ID: 156 / Session SP 02: 4
Short paper (full-text) | 15-minute presentation with a 5-minute Q&A
Keywords: LGBTQ+ fiction, automated indexing, ChatGPT

Efforts to automate cataloging in libraries have progressed significantly, with AI tools like ChatGPT emerging as potential aids. However, automating the subject indexing of LGBTQ+ fiction poses unique challenges. Traditional fiction indexing often overlooks specific themes and characters sought by users, particularly those related to LGBTQ+ identity and issues. Generative Pre-trained Transformers (GPTs) like ChatGPT offer promise in producing detailed subject terms but face biases and inaccuracies. This study explores ChatGPT's efficacy in generating subject index terms for LGBTQ+ fiction by comparing AI-generated terms with those assigned by professional information specialists in the Queerlit database. The Queerlit database, which uses the QLIT thesaurus for LGBTQ+ terms and general Swedish controlled vocabularies, provides a gold standard for this comparison. Using a sample of 20 full-text works and 20 metadata records from the Queerlit database, ChatGPT was tasked with generating subject index terms. The evaluation revealed that ChatGPT struggled to identify any LGBTQ+ themes, often producing broader and irrelevant terms, even when the index terms were given as input in the metadata. The precision and recall scores were low, highlighting AI's limitations in this context. The study underscores the need for careful evaluation of AI tools in library and information science and profession, particularly for indexing fiction and minority representation. Future research should involve collaboration with both information and subject experts to examine the potential of automatically generated terms that were not previously assigned, as well as to examine the possibility of refining automated indexing methods and address inherent biases in AI models.

Koraljka Golub[1], Marianne Ping Huang[2], Isto Huvila[3], Ahmad Kamal[1], Jonas Ingvarsson[4], Olle Sköld[3], Mikko Tolonen[5]
[1]Linnaeus University, Sweden; [2]Aarhus University, Denmark; [3]Uppsala University, Sweden; [4]University of Gothenburg, Sweden; [5]University of Helsinki, Finland

Workshop on Digital Humanities and Social Sciences/Cultural Heritage (DHSS/DHCH) in Higher Education

ID: 134
Half-day conference-themed workshop
Keywords: University programs, courses, open educational resources

Introduction

In recent years, there has been a significant rise in the development of Digital Humanities and Social Sciences (DHSS) as well as Digital Humanities and Cultural Heritage (DHCH) programs across the Nordic, Baltic, and other regions. These programs reflect a growing interest in integrating digital methods into traditional humanities, social sciences, and cultural heritage fields, aiming to prepare students for the changing landscape of academic research and societal needs. Notably, the University of Gothenburg launched a Master's in Digital Humanities in 2017, followed by similar initiatives at Uppsala University in 2019 and Linnaeus University in 2020. The University of Zadar in Croatia has in collaboration with Linnaeus University started a series of BALADRIA summer schools in Digital Humanities with the first introduced in 2019, while the University of Helsinki has long offered modules in Digital Humanities and is now starting a master's program that combines both digital humanities and social sciences. Many other universities have started to offer digital methods courses, either as part of new or existing programs, or as components integrated into broader academic courses. As these programs proliferate, they come with a wide variety of approaches, areas of focus, and teaching materials. Additionally, instructors and program administrators face numerous pedagogical and infrastructural challenges. This makes it crucial to develop platforms where educators, administrators, and researchers can share insights, collaborate on solutions, and create more sustainable and effective learning environments. The proposed workshop aims to address these issues by providing a space for discussion, knowledge exchange, and future collaboration in DHSS/DHCH education. Furthermore, as the eighth annual Higher Education workshop to be held, the 2025 workshop will revisit topics and discussions from previous years, "taking stock" to assess the progress that had been made, the lessons learned, and the capacity that still needs developing in DH higher education.

Importance of the Workshop Topic

Developing and maintaining DHSS/DHCH programs presents several challenges: many institutions face a lack of standardized pedagogical approaches, varying levels of infrastructure support, and the need for ongoing curriculum development to match the rapid pace of technological change and field development. This workshop offers educators and administrators an opportunity to come together and address these challenges through a collective sharing of experiences, teaching methods, and insights. It encourages collaboration to ensure that students receive a rich, interdisciplinary education that prepares them for the complexities of both academic and professional environments. With this aim in mind, the current workshop will include for the first time a systematic "checks-in" on DH education initiatives by partners from previous years. By reviewing the goals, practices, results, and challenges presented thus far in the seven-year run of the workshop, this year's workshop will be especially important for providing a more comprehensive insight of educational programmes within DHSS/DHCH, and exploring translating these insights into new opportunities for collaboration and advancement.

The importance of this workshop lies in its focus on building sustainable infrastructure for future collaboration. By exploring opportunities for joint educational programs, student exchanges, and pedagogy seminars, the workshop paves the way for more coordinated and impactful DHSS/DHCH initiatives across different universities and regions. The long-term goal is to build capacity for cross-institutional cooperation, ensuring that the benefits of interdisciplinary and international learning in digital humanities and social sciences can be fully realized.

Target Audience

The primary audience for this workshop consists of course instructors, program managers, and educational researchers working within DHSS/DHCH programs, courses and online educational materials. These individuals are directly responsible for shaping and delivering curricula, and thus, they are the most engaged in addressing the pedagogical and infrastructural challenges that arise in DHSS/DHCH education. This audience is also most likely to benefit from the exchange of ideas and resources that the workshop aims to facilitate.

In addition to instructors and program managers, the workshop also targets researchers focused on education in digital humanities, social sciences, and cultural heritage. These participants bring valuable perspectives on pedagogical theory, curriculum development, and the integration of digital tools into teaching and learning practices. By involving educational researchers, the workshop ensures a well-rounded discussion that considers both practical teaching strategies and the theoretical underpinnings of DHSS/DHCH education.

The workshop is open to educators from a variety of institutions, including those who are already involved in well-established DHSS/DHCH programs as well as those who are developing or planning to launch similar initiatives. The workshop also invites proposals from those working with online e-learning platforms in the field (e.g., DARIAH Teach, DARIAH Campus, The Programming Historian, Ranke2 etc.). This diverse audience will help create a dynamic exchange of ideas, allowing participants to learn from both the successes and challenges encountered by their peers.

Expected Outcomes

The workshop is designed to foster collaborative discussions and produce concrete outcomes that will benefit the broader DHSS/DHCH community. Some of the expected outcomes include:

1. **Exchange of Pedagogical Approaches and Experiences**: One of the key outcomes of this workshop will be the sharing of teaching methods, tools, platforms, and evaluation strategies used in DHSS/DHCH programs. Participants will have the opportunity to present their experiences with project-based learning, problem-based learning, and interdisciplinary collaboration, which can serve as models for other institutions looking to enhance their curricula.

2. **Development of Collaborative Initiatives**: The workshop will set the groundwork for future collaboration among universities offering DHSS/DHCH programs. This could include initiatives such as student exchanges, joint teaching programs, and regular

pedagogy seminars that allow educators to continue sharing best practices and innovations in the field. The goal is to create a more integrated and collaborative DHSS/DHCH educational landscape.

3. **Infrastructure Building for DHSS/DHCH Education**: Another important outcome will be the exploration of infrastructural needs and solutions to support DHSS/DHCH education. This includes discussions on the resources required to deliver digital methods courses effectively, from computational tools and software to access to online learning resources, datasets and digital archives. Participants will explore how institutions can share resources to overcome infrastructural limitations and support one another in providing high-quality education.

4. **Capacity Building for Student Employability**: A central concern in DHSS/DHCH education is ensuring that students are equipped with the skills needed to succeed in a rapidly changing job market. The workshop will explore strategies for improving student employability by integrating digital methods, critical thinking, and interdisciplinary learning into curricula. This could involve collaborations with industry partners, the development of internship programs, or the inclusion of real-world projects in DHSS/DHCH courses.

5. **Creation of a Community of Practice**: One of the long-term goals of the workshop is to build a sustainable community of practice among educators in the field of digital humanities, social sciences, and cultural heritage. This community can serve as a platform for ongoing collaboration, resource sharing, and mutual support as DHSS/DHCH programs continue to evolve. By creating a network of educators who are invested in the success of these programs, the workshop ensures that the discussions and initiatives begun here will have a lasting impact on the field.

Workshop Structure

The workshop will be structured to maximize participant engagement and facilitate productive discussions. It will be divided into five sessions:

1. **Session 1: Welcome and Introductions (15 minutes)**

This opening session will provide participants with an overview of the workshop's goals and structure. It will also give attendees an opportunity to introduce themselves and their areas of interest, setting the stage for collaborative discussions.

2. **Session 2: Presentation and Discussion of Submitted Papers (120 minutes)**

In this session, participants will present papers submitted through an Open Call. Each presentation will last approximately 10 minutes, followed by a short discussion. These presentations will address various topics related to DHSS/DHCH education, such as interdisciplinary cooperation, project-based learning, program development, and student employability. The session will provide a platform for sharing practical experiences and innovative teaching strategies.

3. **Session 3: Taking stock (60 minutes)**

This session involves workshop coordinators and participants from previous workshops to specifically revisit the topics presented in the past. The goal of this session is to follow-up on initiatives presented in earlier sessions, distill the lessons-learned, identify ongoing or new challenges, and grasp the trends in topics over the years to offer a wider perspective to inform future planning in DHSS/DHCH programmes.

4. **Session 4: Educational Hack-a-thon (45 minutes)**

In order to begin addressing the issues raised in from the previous session, participants will be asked to work in teams to brainstorm educational initiatives which can potentially address concerns or exploit opportunities.

5. **Session 5: Directed Discussion (30 minutes)**

This final session will build on the themes and ideas that emerged from the previous presentations. Participants will engage in a directed discussion aimed at identifying key areas for future collaboration, including potential joint programs, infrastructure-sharing initiatives, and student exchange opportunities. The goal of this session is to translate the insights gained from the workshop into actionable plans for the future.

Vojko Gorjanc[1,2]
[1]University of Ljubljana, Faculty of Arts; [2]Institute of Contemporary History, Ljubljana

Exploring Digital Frontiers: Mastering Linguistics and Humanities Collaboration

ID: 281 / WS09: 1
Workshop on Digital Humanities and Social Sciences/Cultural Heritage (DHSS/DHCH) in Higher Education
Keywords: **digital linguistics, digiral humanities, MA program, competences, interdisciplinary skills**

This presentation examines the collaborative framework of the joint MA program in Digital Linguistics, an interdisciplinary initiative co-developed by the University of Ljubljana, Masaryk University in Brno, and the University of Zagreb (https://digiling.university/). The program's design focuses on fostering competencies that extend beyond digital linguistics, incorporating methodologies from digital humanities to offer a holistic, future-oriented academic experience. The curriculum is structured around three core competency clusters: linguistics, information technologies, and social sciences, which together equip students with the skills necessary for interdisciplinary research and professional practice.

A significant aspect of the presentation is the collaborative model implemented during the study segment at the University of Ljubljana, where students work closely with the Digital Humanities research group and the DARIAH-SI research infrastructure. This segment highlights the deliberate integration of pedagogical and research activities, aiming to cultivate both interdisciplinary skills and tangible, real-world outcomes. As part of the program, students are introduced to ongoing research initiatives and European research infrastructures, including CLARIN.SI and DARIAH-SI, through workshops and collaborative opportunities. These interactions not only immerse students in cutting-edge projects but also connect them with the principles of open science, which are increasingly shaping the academic landscape, particularly in Slovenia, where open data practices in the humanities are still developing. One noteworthy outcome of this collaboration is the DIHUR (Digital Humanities Research) podcast series, a student-driven initiative guided by faculty and researchers (https://www.youtube.com/@DigitalnaHumanistika). These podcasts explore different topics in digital humanities and feature interviews with experts, bridging the gap between scholarly research and public discourse. By participating in projects like these, students gain hands-on experience in content creation, research dissemination, and public engagement, all within a professional academic environment.

This presentation underscores the program's innovative approach to integrating education and research. By fostering active collaboration between students, researchers, and research infrastructures, the MA program not only prepares students for careers in academia and beyond but also demonstrates how interdisciplinary education can drive societal and academic impact.

Edward Joseph Gray[1,2], **Sanita Reinsone**[3], **Koraljka Golub**[4,5], **Mikko Tolonen**[4,5], **Johanna Lilja**[4,5], **Inés Matres**[4,5], **Eiríkur Smári Sigurðarson**[6], **Olga Holownia**[6], **Mari Väina**[7]

[1]DARIAH-EU, France; [2]IR* Huma-Num (CNRS); [3]Institute of Literature, Folklore and Art of the University of Latvia; [4]University of Helsinki; [5]DARIAH-FI; [6]Centre for Digital Humanities and Arts, University of Iceland; [7]Estonian Literary Museum

How to Structure and Organize a National Digital Humanities Research Infrastructure: Realizing the Digital Dreams of Tomorrow

ID: 215

Half-day tutorial

Keywords: Research Infrastructure, Digital Turn, Research Data Management, Community Building, Science Policy

In honor of DHNB2025's theme of Digital Dreams and Practices, we propose a workshop on research infrastructure - in essence, how to put into practice the realization of these digital dreams.

Research Infrastructure is a necessity for effective and productive research outcomes, and indeed, it has been recognized as such by the European Commission, who invented the ESFRI Scheme to ensure a strategically-planned, harmonized European Research Area. Nevertheless, while it is one thing to have a continent-wide plan, it is another matter entirely as to how to implement this plan and operate it on a daily basis on a national level. Further complicating matters are different national and regional research traditions, and the need to respond to the needs of the research community - both in terms of technical solutions, and in providing an effective (human) interface to counsel and aid these users in the uptake of solutions. Yet another challenge is ensuring access to data, both in terms of the technical aspects (managing versioning, provenance, long-term preservation, and providing for interoperability) and juridical questions that are inherent to responsible data management.

This workshop will discuss and lay out what the presenters - who have deep experience in the construction of a research infrastructure - feel is necessary to build an effective national research infrastructure for the digital humanities community. It will involve presenters from Sweden, Finland, Iceland, and Latvia and is open to all members of the DHNB community looking to build a national research infrastructure for social sciences and humanities.

While not exclusively dedicated to DARIAH, the European Research Infrastructure Consortium dedicated to the arts and humanities, this workshop will focus in large part on how DARIAH's experiences, both on the European and national level, can be helpful for those countries which wish to launch their own national research infrastructure. Indeed, with ten years of existence and 23 Member countries (as of October 2024), DARIAH represents a successful model to empower research communities with digital methods to create, connect and share knowledge about culture and society. DARIAH's 4 Strategic Pillars, comprising (I) a Marketplace of reusable tools, services, data, and knowledge; (II) Education and Training, (III) Transnational and Transdisciplinary Working Groups, and (IV) Policy and Foresight represent essential components of a research infrastructure - ensuring that the infrastructure provides something useful to researchers, that it ensures its communities are aware of the most recent practices, that it integrates feedback from their needs, and keeps an eye on policy developments to both influence them and best prepare its community for what comes next.

After an introductory talk on what research infrastructure is, and how they fit into the strategic framework of research policy, the workshop will then proceed to presentations that introduce four use-cases on how national research infrastructures have developed, what challenges they faced, and how they build community. Then, we will have a discussion amongst participants to see where they are in the process of building their national research infrastructure, and what their next steps are in realizing the digital dreams of tomorrow.

Target Audience:

The targeted audience of this workshop includes both researchers (though typically those that lead institutions, labs, projects, or are otherwise inclined to think in an infrastructural manner), research engineers, and policymakers from ministries. Our wish is to bring together representatives from all DHNB countries that are building, or beginning to build, their national research infrastructure and give them a forum to share their concerns and receive feedback from our workshop leaders, and the other attendees.

We cannot count on their attendance at this early date, but if accepted, we wish to have participation from Ministry officials, particularly in Estonia and other Baltic nations - to both have their insight on the question of research infrastructure, but also for them to learn about the impact that a national research infrastructure modeled on DARIAH can provide.

Expected Outcomes:

The goal is to bring together the different stakeholders (users, providers, funders) of the DHNB community that are involved in the construction of a research infrastructure, and give them a forum to exchange, learn, and begin to construct together a plan of action. There is a great diversity of national research infrastructure in the DHNB community, with countries with full-fledged national infrastructures, in the process of building these RIs, and those that are at the very beginning, assembling community and political support. This event, we hope, will help the community share notes, exchange best practices, and start to think about what a research infrastructure ought to look like in the Nordic & Baltic region and beyond.

The hope is that this workshop will both jumpstart countries - particularly the Baltic countries - that are in the process of building a national research infrastructure in digital humanities, and give a forum for currently-established national research infrastructures, and in particular those in the Nordic and Baltic region, to exchange best practices and contacts. Particularly, we hope to profit from the host country's effort to join DARIAH, which is now on the Estonian Research Infrastructure Roadmap, to give them, and their Latvian and Lithuanian partners, a leg up in the foundation of their national research infrastructures. As well, this workshop will continue the discussions we had in Reykjavik at DHNB 2024, helping to concretize the work being done in Iceland, Finland, Sweden, and Norway.

Proposed Schedule

1. Opening Address: Services, Strategy, and Scholars: The Role of Research Infrastructure in Enabling Research
2. DARIAH-EU as a Structuring Mechanism for Building National Research Infrastructure
    1. Discussion - What do you expect from a Research Infrastructure?

3. *(Potential Space for word from Ministerial representatives)*
4. National Use Cases for Building Research Infrastructure
    1. Latvia
    2. Iceland
    3. Sweden
    4. Finland
5. Panel Discussion from National Research Infrastructure Use Cases
6. Breakout Brainstorming Sessions What does your ideal national Research Infrastructure Look like? What are the major challenges? How good are your contacts with the Ministry? What help would you appreciate, either from other national RIs or ERICs like DARIAH ?
7. Restitution of Breakout-Group Findings
8. Conclusions & Paths Forward

Hinrik Hafsteinsson[1,2], **Steinþór Steingrímsson**[1]
[1]The Árni Magnússon Institute for Icelandic Studies; [2]University of Iceland

Natural Language Processing for Everyone: The Case for a Centralized Icelandic NLP Platform

Thursday, 06/Mar/2025 2:00pm - 2:20pm
ID: 222 / Session SP 07: 1
Short paper (full-text) | 15-minute presentation with a 5-minute Q&A

Keywords: Icelandic, Language Technology, NLP, Accessibility, API

This paper presents the Árni Magnússon Institute's NLP platform, a website where the functionality of important language technology tools is made accessible to the general public, both through a user interface and a standardized API. The platform is presented as a solution to the problem that even though a specific language technology tool exists, it does not go without saying that anyone can use it: A certain technical know-how is almost always a prerequisite for being able to use these tools. The platform is presented with the following NLP solutions integrated: Tokenization, PoS-tagging, Lemmatization and Hyphenation, with more tools becoming available pending future development. The platform is now made available to the public, with the caveats of it being in active development and undergoing regular changes.

Gert Foget Hansen
University of Copenhagen, Denmark

Running Whisper in Praat – Simplifying local setup of Whisper, expanding output options

ID: 236 / Poster Session 2: 13
Poster and demo (abstract) with accompanying a 1-minute lightning talk
Keywords: Praat, Whisper, Speech-to-text, STT, transcription

One of the challenges with audio recordings is that it can be extremely difficult and time consuming to navigate any substantial amount of material if no detailed annotations or transcriptions exist. Automatic transcription tools such as Whisper can provide rough transcriptions of speech at little cost. In spite of the varying accuracy, such transcriptions can facilitate navigation in a collection of audio material that would otherwise be much less accessible.

While useful, the immediate output from tools such as Whisper has certain limitations. For instance: Whisper seems best suited to monologue material and makes no attempt to distinguish between multiple speakers. As Whisper was originally designed to provide subtitles for video, the length of utterances often seems to be determined by what amount of text can be shown on screen at the same time rather than to be in accord with punctuation, pauses or answer question sequences.

In order to expand the usefulness of Whisper, and to make it easier to install and use for non-expert users, the Whisper_in_Praat script package was created. It is based on a derivative of OpenAI's Whisper that comes as a compiled executable. The enhancements to the output come by post processing the output from Whisper and if needed also by preprocessing the input audio signal.

The Whisper_in_Praat script package has been created to provide:

- A simple installation process
- Local processing of media files
- A basic graphical user interface
- Options to generate more advanced types of output

Post-processing of the output from Whisper – to some extent in combination with further audio analysis – provides additional output formats:

- TextGrid (a tier-based format native to Praat and compatible with other transcription and annotation software such as ELAN) at utterance and word level
- HTML with links to audio

as well as other enhancements:

- Removal of punctuation and uppercase letters (as is the standard practice for transcriptions in most spoken language research)
- Improved delimitation of utterances
- Distinguishing two or three speakers in two-channel recordings, provided the speakers are recorded with a suitably high channel separation

The accompanying preprocessing scripts are used to identify sound from individual speakers in two channel recordings of two speakers. By muting the cross talk the preprocessed files allows Whisper to transcribe two speakers individually.

Bibliography

The Praat script makes use of Whisper-faster executables from
https://github.com/Purfview

Those executables are based on faster-whisper that is available from
https://github.com/SYSTRAN/faster-whisper

Those are in turn is based on OpenAI 's Whisper:
https://openai.com/research/whisper

Latest public iteration of Whisper_in_Praat is available at
http://dx.doi.org/10.13140/RG.2.2.24093.93925

Louise Brix Pilegaard Hansen[1], Jan Kostkan[1], Roberta Rocca[2], Kristoffer Nielbo[1]
[1]Center for Humanities Computing, Aarhus University, Denmark; [2]Interacting Minds Centre, Aarhus University, Denmark

Multimodal pre-training of vision models yields better embeddings for visual art

Thursday, 06/Mar/2025 9:30am - 10:00am
ID: 132 / Session LP 06: 2
Long paper (full-text) | 20-minute presentation with a 10-minute Q&A
Keywords: Multimodal Models, Visual Art Analysis, Image Embeddings

Deep pre-trained vision models provide automated ways of analyzing large digitized corpora of visual art. Central to the success of these models is their ability to extract rich embeddings for downstream tasks such as style classification or painting retrieval. Recent results suggest that multimodal models trained on a combination of visual and linguistic input yield semantically enhanced and higher-quality representations of images compared to unimodal vision models. While these multimodal models seem extremely promising for the computational study of visual art, where semantic knowledge might be relevant to produce informative representations of artworks, research benchmarking multimodal models for feature extraction in the art domain is limited, and their potential for representing visual art remains to be explored. This paper aims to fill this gap, and it compares the representational abilities of seven unimodal and multimodal state-of-the-art pre-trained vision models by employing their embeddings in three domain-specific downstream tasks: genre classification, style classification, and artist classification in the WikiArt dataset. Results reveal that multimodal models perform best as feature extractors for artworks as opposed to unimodal models. We hypothesize that pre-training on natural language descriptions provides multimodal models with enhanced abilities to infer global semantic representations of an image, which is beneficial to identifying key characteristics of artworks such as their genre, author, and style.

Lauri Matias Heinonen
University of Bamberg, Germany

Automating the recognition of semi-structured tax data: Insights from applying HTR to the Schenkenschans customs registers (1630-1810)

ID: 189 / Poster Session 2: 5
Poster and demo (abstract) with accompanying a 1-minute lightning talk
Keywords: HTR, semi-structured data, workflow, tax data, automation

The DFG-funded project „The ‘invisible carriers’ on the Rhine, 1630-1810. Observing the early modern economy through the lens of freight transport practices“ (Proj.Nr. 527636627) develops an HTR-based pipeline for the automated transcription of the customs registers of Schenkenschans (1630-1810), a fortress and a customs station on the Rhine at the border between Germany and the Netherlands, to study early modern international trade in the Rhine region.

The contribution discusses the degree to which the recognition with HTR tools of the semi-structured data of the Schenkenschans customs registers can be automated. I understand semi-structured data as the combination of more or less formalized text entries on the one hand, and columns that register the taxes, on the other. The tax columns are often explicit, with vertical lines separating columns, but sometimes they are merely implied by the use of dashes and dots.

In this paper, I focus on the recognition and revision of tax data in the customs entries. The first version of the transcription was created with text recognition models, which we trained to capture and imitate the partly columnar structure of the registers. I compare the recognition results with a first revision of the tax columns to assess the quality of our tailed-made recognition models. The comparison and quality control helps me to identify where, when and how manual processing steps for this kind of data are required. My analysis helps me to highlight the importance of segmentation, model training and manual quality control of specific parts of the transcription as crucial for the implementation of workflows that aim at capturing semi-structured or partly columnar data.

Some contributions like Hodel et al. (2021) note that automated text recognition tools often do not achieve a good Character Error Rate (CER), e.g. below 5%. Thus, they suggest that high quality transcription and the whole data pipeline involves a decent amount of manual work. Yet other contributions suggest that automated processing can achieve quite good results. Developing methods for transciption, Capurro et al. (2023) use a correction algorithm to improve an initial automated transcription of multilingual archives of a Dutch glass artist and compare the results of both steps. They are able to achieve a significant reduction of the manual labour necessary in the process.

While addressing ways to automate data collection and processing, I also want to discuss phases where some manual work is required or even preferred. Automated data processing should find a balance between speeding up data processing through automation and ensuring quality through rigorous control mechanisms that may involve manual labour.

Bibliography

Capurro, Carlotta; Provatorova, Vera; Dupre, Sven; Hendriksen, Marieke & Kanoulas, Evangelos (2023). Chasing the Model. Experimenting With Training a Neural Network to Recognise Text in a Multi-Language and Multi-Authored Handwritten Document Collection. Abstract published in DHNB2023 Book of Abstracts.

Hodel, T., Schoch, D., Schneider, C., & Purcell, J. (2021). General Models for Handwritten Text Recognition: Feasibility and State-of-the Art. German Kurrent as an Example. Journal of Open Humanities Data, 7: 13, pp. 1–10. DOI: https://doi.org/10.5334/johd.46

Olga Holownia[1], **Gustavo Candela**[2], **Helena Byrne**[3], **Jon Carlstedt Tønnessen**[4], **Anders Klindt Myrvoll**[5], **Sophie Ham**[6], **Steven Claeyssens**[6]

[1]IIPC, United States of America; [2]University of Alicante, Spain; [3]British Library, UK; [4]National Library of Norway, Norway; [5]Royal Danish Library, Denmark; [6]KB, National Library of the Netherlands

Web Archive Collections as Data

ID: 207

Half-day conference-themed workshop

Keywords: collections as data, researcher engagement, Jupyter Notebooks, corpora

GLAM (Galleries, Libraries, Archives and Museums) have started to make available their digital collections suitable for computational use following the Collections as Data principles[1]. The International GLAM Labs Community[2] has explored innovative and creative ways to publish and reuse the content provided by cultural heritage institutions. As part of their work, and as a collaborative-led effort, a checklist[3] was defined and focused on the publication of collections as data. The checklist provides a set of steps that can be used for creating and evaluating digital collections suitable for computational use. While web archiving institutions and initiatives have been providing access to their collections - ranging from sharing seed lists to derivatives to “cleaned” WARC files - there is currently no standardised checklist to prepare those collections for researchers.

This workshop aims to involve researchers and web archive practitioners in reevaluating whether the GLAM Labs checklist can be adapted for web archive collections. The first part of the workshop will introduce the GLAM checklist, followed by four use cases that show how different web archiving teams have been working with their institutions’ Labs to prepare data packages and corpora for researchers. For the second part of the workshop, we want to issue a separate call for use cases and have researchers present examples of their use of web archive collections and discuss their workflows. In the final part, we want to involve the audience in identifying the main challenges to implementing the GLAM checklist and determining which steps require modifications so that it can be used successfully for web archive collections.

First use case

The UK Web Archive has recently started to publish the metadata to some of our inactive curated collections as data. This project developed new workflows by using the Datasheets for Datasets framework to provide provenance information on the individual collections that were published as data.

In this presentation, we will highlight how participants can:

- Use Datasheets for Datasets to describe their collections.
- Potential research uses for the data sets that were published.
- Gain insights from the lessons learnt phase of the project.

Second use case

The National Library of Norway (NLN) has recently launched its first ‘Web News Collection’, making more than 1.5 million texts from 268 news websites openly available for computational analysis via API. The objective is to facilitate computational text analysis of news content from the web, both with computational notebooks and user-friendly web apps [4].

This presentation will:

- Briefly explain the ‘warc2corpus’ pipeline for extracting natural language from Web ARChive (WARC) files and preparing text data for computational analysis,[5]
- Show the metadata schema and overall statistics for the collection,
- Demonstrate how students, scholars and others can tailor their own corpora and perform various forms of ‘distant reading’,
- Reflect on how the ‘Web News Collection’ aligns with the FAIR principles while also taking immaterial rights into account.

Third use case

The Royal Danish Library has recently launched a service for everyone to free text search in the entire Danish web archive using Smurf - N-gram visualisation[6]. The free text search will search in HTML-pages for each year, and the number of results found will be compared to the total number of HTML from that year. The total number of HTML pages in the archive is currently more than 20 billion pages (20.000.000.000), spanning from 1995 to the present. Due to legal reasons, only one or two words can be used in the search, and no special characters are allowed. Normally, you can only gain access to the Danish web archive upon application if you are a researcher or PhD student affiliated with a Danish research institution. So this is a possibility for everyone to gain insights and see trends from an otherwise very restrictive and closed archive.

This presentation will:

- Show how to use the NGRAM search for feasibility studies for researchers before applying for access to the archive, as well as for the general public who would like to have a sneak peek into the archive.
- Look into the needs, practicalities, and possibilities of sharing corporate data or metadata with researchers and the general public.
- Explore particular situations concerning data availability that can be useful to extend the scope of the GLAM Labs checklist as well as provide additional examples of application.

Fourth use case

The KB, National Library of the Netherlands, has been refining its strategy for Collections as Data, with a stronger focus on implementing FAIR principles, documenting data provenance, and transparency of selection workflows. A critical question in this effort is how elements of the GLAM Labs checklist can be adapted for use at different levels: collection-specific (e.g., web collections) versus a more institution-wide, generic approach.

This presentation will:

- Showcase the current capabilities of KB in offering web collections as data, along with their descriptions.
- Explore the adaptability of the GLAM Labs checklist for both collection-specific applications and broader institutional use
- Offer a glance at our future plans, concerning collections as data and inviting the participants to provide feedback.

Expected outcomes:

- Understanding the steps involved in preparing digitised and born-digital collections for publication
- Understanding the challenges involved in preparing different types of web archive collections for publication and/or sharing with researchers
- Creating a draft checklist for publishing web archives collections as data

Bibliography

[1] Padilla, T. (2017). “On a Collections as Data Imperative”. UC Santa Barbara. pp. 1–8;

[2] https://glamlabs.io/

[3] Candela, G. et al. (2023), "A checklist to publish collections as data in GLAM institutions", Global Knowledge, Memory and Communication. https://doi.org/10.1108/GKMC-06-2023-0195

[4] Tønnessen, J. (2024). “Web News Corpus”. National Library of Norway. https://www.nb.no/en/collection/web-archive/research/web-news-corpus/; “Apper fra DH-lab”. National Library of Norway. https://dh.nb.no/apps/.

[5] Bremnes T,. Birkenes M., Tønnessen J. (2024). “corpus-build”. GitHub. National Library of Norway. https://github.com/nlnwa/corpus-build; Birkenes M., Johnsen, L., Kåsen, A. (2023). “NB DH-LAB: a corpus infrastructure for social sciences and humanities computing.” CLARIN Annual Conference Proceedings.

[6] https://www.kb.dk/en/find-materials/collections/netarkivet linking to https://labs.statsbiblioteket.dk/netarchive/ngram/#/

Agata Hołobut[2], Maciej Rapacz[1], Miłosz Stelmach[2]
[1]AGH University of Kraków, Poland; [2]Jagiellonian University in Kraków

Untranslated film titles in cross-national distribution: Metadata analysis

Thursday, 06/Mar/2025 3:00pm - 3:20pm
ID: 220 / Session SP 05: 4
Short paper (abstract) | 15-minute presentation with a 5-minute Q&A
Keywords: corpus linguistics, natural language processing, film studies, metadata analysis

Film titles are paratexts (Genette 1997), which may perform a number of functions: (a) distinctive, i.e., helping differentiate the production from others; (b) metatextual, i.e., presenting themselves as film titles; (c) phatic, i.e., grabbing people's attention; (d) referential, i.e., revealing information about the plot and genre of the production; (e) expressive, i.e., conveying the filmmakers'/production studio's attitudes towards the movie and the world at large; and (f) appellative, i.e., persuading potential viewers to watch the movie (Nord 1995). It is the most important and the cheapest means of promotion, which determines the recognizability of the cinematic production and may influence its box-office success (Bae and Kim 2019). This is why the characteristics of "effective" film titles have been investigated to date primarily by marketing and business scholars (Sood and Drèze 2006; Bae and Kim 2019; Chung and Eoh 2019; Xiao and Cheng 2021).

A phenomenon less explored, yet equally important, concerns translation of original titles for the purposes of international film distribution. This problem has been mostly explored by translation scholars working on small bilingual corpora of film titles, focusing mostly on most recent trends. They have distinguished various strategies and procedures of title translation and conducted case studies for specific countries (e.g. Gabrić et al 2022, Pastor 2011, Shokri 2014, Kim 2017, Fakharzadeh 2022).

Our research aims to enrich this body of research by accomplishing a quantitative analysis of a large corpus of film titles translated into multiple languages. It is based on a collaboration of a team comprising of a computer scientist, film scholar and translation scholar and expands a series of explorations we have already conducted into global circulation of cinematic productions, based on film metadata analysis. Film metadata, i.e., the verbal and numerical cultural information preserved in film databases, have been used by scholars to date in multiple contexts, for example to retrace the changing film production practices in terms of growing crews (e.g. Tinits and Sobchuk, 2020); investigate national and institutional framing of the cinematic production (cf. e.g. Van Beek and Willems, 2022), follow distribution circuits (Zemaityte et al. 2023) and spot cultural trends (Viacom 2017).

In our research we have so far explored the evolution of Hollywood film remakes and sequels (Stelmach, Hołobut, Rybicki 2022; Hołobut, Rybicki, Stelmach 2024) and traced the patterns in film-title translation around the world (Hołobut, Rapacz, Stelmach 2024a and 2024b). This time we build on the findings of the latter project, which revealed a steady global increase in the annual share of feature films distributed internationally under their original titles, un-adapted to the recipient languages and cultures, from around 20% in the 1950s to more than 30% in the 2020s.

Interested by these findings, in our current research we address the following research questions and areas:

1. Geographical – How has the ratio of untranslated titles grown in specific regions and subregions of film distribution? Are there regions in which this growth has not been observed? Are there regions where this growth is particularly substantial? How to explain these tendencies?
2. Cultural – Which languages do the transferred titles come from? Is their steadily rising presence connected with the popularization of English as a lingua franca?
3. Cinematic – Do arthouse and mainstream productions behave in a similar way regarding the growth of untranslated titles? Does the presence of "untranslated" titles depend on film genre?
4. Linguistic – What language features do the "untranslated" titles share? How complex are they in terms of complexity and length (measured in terms of number of characters and words, as well as vocabulary richness)?

To explore these questions, we analyze multiple subsets of a corpus containing feature film titles released between 1950 and 2023. This corpus was created through compilation of datasets from the Internet Movie Database (IMDb) and spans over seven decades of cinematic releases. The subsets include:

1. The complete dataset, comprising 453,517 films and 2,026,710 localized title tokens. We analyze all deciles of the data based on the number of title variants, aiming to assess whether the trends observed in the most popular, and likely best-annotated, productions extend to the broader dataset. For comparison we also use a smaller, finely-tuned datasets comprising of most internationaly successful types of productions:
2. The annual top fifty productions distributed in the most language versions since 1950, as recorded by IMDb. This subset consists of 3,700 films and 184,459 localized title tokens, largely distributed by major Hollywood studios.
3. All localized title variants (referred to as "AKAs" - "also known as") of feature films that have been part of the Official Selection at the Cannes Festival since 1950. These films, predominantly art-house productions, are very often distributed by non-Hollywood companies and cater to a global cinephile audience. This subset includes 3,130 films, with a total of 63,754 localized title tokens.

Each record in our dataset contains the number of attributes, that can be computated, represented graphically, analysed etc. They include original movie title, year of release, localised movie title (the title under which the movie was distributed in a given region), region of distribution, genre(s) of the movie, country of origin.

Our preliminary research confirms the predominance of English as the language most frequently used in global circulation of films with unchanged titles. It also points to the regional variation in the trend, with most steep rise of share of untranslated titles in Western Europe, Middle East and Africa along less pronounced shifts in other regions. Less variation can be observed among the various genres of productions, with the exception of historical films, where tendency to use a non-translated title is much stronger.

Bibliography

Bae, G. and Kim H. (2019). “The impact of movie titles on box office success”. Journal of Business Research 103: 100 – 109.

Chung, J. and Eoh, J. (2019). “Naming strategies as a tool for communication: Application to movie titles”. International Journal of Advertising 38 (8): 1139–1152.

Fakharzadeh, M. (2022). “A sociological approach to official and non-official audiovisual translators’ practice in Iran: The case of movie title translation.” Journal of Intercultural Communication Research, 51(4): 430-449.

Gabrić, P., Brajković, I. Kelčec Ključarić, D. and Bezuh, J. 2022. “A comparative and diachronic analysis of film title translations and appellative effect transfer into Croatian and German”. doi:10.21203/rs.3.rs-1220241/v1.

Genette, G. (1997). Paratexts: Thresholds of Interpretation, Cambridge University Press. Cambridge.

Hołobut, A., Rapacz, M. and Stelmach, M. (2024a). “Working titles. Computational analysis of film titling practices: A Polish case study.” Kwartalnik filmowy 127: 133-162.

Hołobut, A., Rapacz, M. and Stelmach, M. (2024b). “Translating film titles: a qualitative and quantitative approach”. The Palgrave Handbook of Multilingualism and Language Varieties on Screen, ed. I. Ranzatto and P. Zabalbeascoa. Cham: Palgrave Macmillan, 618-638.

Hołobut, A., Rybicki, J. and Stelmach, M. (2024). “A statistical approach to Hollywood remake and sequel metadata”. Digital Scholarship in the Humanities 39 (2): 556-574.

Kim, W. (2017). “Lost in translation. (Mis)translation of foreign film titles in Korea”. Babel 63 (5): 729-745.

Nord, C. (1995). “Text-functions in translation: Titles and headings as a case in point 1’. Target 7 (2): 261–84.

Pastor, L. (2011). “Las leyes de la atracción en los títulos de las películas de cine. Un caso de análisis de la vía heurística en la comunicación de masas”. Anàlisi 43: 89-103.

Shokri, S. (2014). “Translating movie titles: Strategies applied on Persian to English cases”. Mediterranean Journal of Social Sciences 5 (20): 2568-2572.

Sood, S., and Drèze, X. (2006). “Brand extensions of experiential goods: Movie sequel evaluations”. Journal of Consumer Research 33 (3): 352–60. doi:10.1086/508520.

Stelmach, M., Hołobut, A. and Rybicki, J. (2022). “Quantifying the remake. A historical survey”. Journal of Adaptation in Screen and Performance 15 (3): 211–226.

Tinits, P. and Sobchuk, O. (2020). “Open-ended cumulative cultural evolution of Hollywood film crews”. Evolutionary Human Sciences 2 (26): 1–15.

Viacom (2017). “Intersections of story and culture”. 50 Years of TV, Film and Culture. http://fantheory.viacom.com/ (accessed 5 October 2024).

Xiao, X., Cheng, Y. and Kim, J. (2021). “Movie title keywords: A text mining and exploratory factor analysis of popular movies in the United States and China”. Journal of Risk and Financial Management 14: 68.

Van Beek, B. and Willems, G. (2022). “Intranational film industries: A quantitative analysis of contemporary Belgian cinema”. French Screen Studies 22 (4): 304–325.

Zemaityte, V., Karjus, A., Rohn, U., Schich, M., and Ibrus, I. (2023). “Quantifying the global film festival circuit: Networks, diversity, and public value creation”. DOI: 10.31235/osf.io/g9w4b (accessed 10 October 2024).

Siska Humlesjö, **Julia Beck**, **Kristin Åkerlund**
University of Gothenburg, Sweden

The Dawit Isaak database of censorship - creating a database of forbidden and censored literature

Thursday, 06/Mar/2025 3:00pm - 3:20pm
ID: 155 / Session SP 06: 4
Short paper (abstract) | 15-minute presentation with a 5-minute Q&A
Keywords: Censorship, Linked open data, Bibliographic database

The goal of this database is to:

Give the general public, schools and journalists an easy way to gather information on censored and banned literature

Create a platform for data about censored and banned literature for researchers.

Background

DIDOC, https://didoc.dh.gu.se/ , The Dawit Isaak database of censorship, is a pilot project and collaboration between the Dawit Isaak Library, The Gothenburg Research Infrastructure for Digital humanities, Swedish PEN, and Lund University. The Dawit Isaak library, named after the imprisoned Swedish Eritrean journalist, is a public library dedicated to collecting banned and censored literature, literature about book bans and censoring of literature. As of today their collection of banned and/or censored literature constitutes of approximately 1200 works. The Swedish PEN is the Swedish part of the international PEN, a worldwide organisation for writers that advocates for freedom of expression. The database was beta-released in September 2024 based on a selection of 160 titles.

General overview

Currently, the resource is divided into:

- Titles
- Explore Bans
- Editorial Content,

Titles: The titles listed are a selection made by the Dawit Isaak Library and Swedish PEN for Banned Books Week Sweden. The bibliographic information is sourced from Sweden's union catalogue, Libris. The goal is for the information to be easy to access and for users to quickly find the works in their local libraries.

Explore Bans: Each ban or censorship is added to the database as its own record of an event, connected to a record of a work. As for now we have focused on a simple search instead of a more advanced search. What we do offer is filtering since the focus is browsing, not searching.

The filters are based on the material currently in the databases, creating a possibility to add parameters if the material in the database grows.

The bans can be filtered by three parameters:

Type of ban, such as School Bans, Book burnings, Internet censorship, Self censoring, or Confiscation;

Reason(s) for the ban, like Blasphemous Content, State security, Explicit content, Agitation or LGBTQI Content;

The location where the ban took place, which can be a country, a state, or a historical place like the Soviet Union. The locations are geo-tagged, allowing us in the future to create visualisations of where these events occurred in the world.

For each event, there is information about when the ban took place, which organisations or individuals were involved, a description of the event, and references to the source of the information. Of course, there's also a link to the related title.

Editorial content: The editorial content consists of as of now four articles, written as teacher's guides for banned books week have been reworked to overviews of banning and censorship in Sweden, Iran and China and of Children's literature.

Ethical considerations

Information about banned literature and persecuted authors is by nature highly sensitive. By relying on information from Swedish PEN and the Dawit Isaak Library, the database can ensure that the information is either already publicly available or published with the author's consent.

Technical background

We chose the open source web publishing platform Omeka-S as the database solution for DIDOC due to the possibility to connect to the semantic web by using Linked Open Data. Standards like Dublin Core are already included or can be added easily. Additional vocabularies and ontologies can be imported or created. Other important factors for our choice of Omeka-S were its large community, user-friendliness and its extensibility with further modules that add functionality like improved batch processes as well as visualisations. Especially interesting in the context of a censorship database, was the ability of Omeka-S to provide reification. This helps to annotate statements with for example a degree of certainty or a source of information. We also added the possibility in the data editor to autofill person and organisation authority files by using the API from the Swedish National Library.

Future development

Our plan is to apply for research funding to expand the database. The vision is to create a database that can incorporate both the Dawit Isaak Library's whole collection of around 1200 works and Beacons of Freedom, an earlier database of banned and censored literature of about 50.000 records, created by the National Library of Norway.

We also want to develop visualisations, for example an interactive map and timelines, to provide users with different ways of accessing the information.

Eero Hyvönen[1,2], **Annastiina Ahola**[1], **Petri Leskinen**[1], **Jouni Tuominen**[2,3]
[1]Aalto University; [2]University of Helsinki (HELDIG); [3]University of Helsinki (HSSH)

Aggregating and Aligning Knowledge Graphs into a Global Service: SampoSampo System for Cross-cultural Data Search, Exploration, and Analysis

Wednesday, 05/Mar/2025 4:00pm - 4:20pm
ID: 168 / Session SP 03: 1
Short paper (abstract) | 15-minute presentation with a 5-minute Q&A
Keywords: linked data, data alignment, semantic portal, data analysis

Abstract

This paper abstract presents an approach and first results of creating a global LOD service and semantic portal, "SampoSampo", based on a network of interlinked Cultural Heritage knowledge graphs (KG) of different application domains. In this way, a more comprehensive global view for searching, exploring, and analyzing the interlinked KGs can be provided than by using local KGs separately. The SampoSampo LOD service can be used as a web service for providing IRI identifiers for aligning new datasets with the national DARIAH-FI research infrastructure on which SampoSampo is based, and for assessing the quality of data in different KGs by comparing their metadata. The portal underway can be used for searching and exploring the local KGs by a single user interface (UI) and for implementing novel application perspectives based on relational search, where semantic “interesting” associations (relations) between entities, such as persons, organizations, and places in the global KG, can be searched for and natural explanations for them be created (explainable AI).

1. Enriching and harmonizing data by linking

One of the great promises of Linked Data, as promoted by the 5th star in Tim Berners-Lee 5-star model[1], is enriching data by linking it to external datasets using IRIs. A requirement for this is that the linked datasets use the same identifiers (IRIs) for the same resources or that IRI alignments across datasets are available. For data models and Knowledge Organization System (KOS) schemas, such as Dublin Core[2], RDF[3], SKOS[4], etc., harmonized use of IRIs is often the case due to standardization efforts, but not so often for data resources representing things of the real world, such as persons, organizations, and places.

This key problem of aligning and interlinking datasets based of different identifier systems has been addressed before in many ways. For example, in the library world, different identifiers are often used for, e.g., authors and places in national collections, and linking systems, such as the Virtual International Authority File service VIAF[5] have been created to mitigate the problem (Hickey and Toves 2014). Linked Open Vocabularies (LOV) is an example of a high-quality catalogue of reusable vocabularies, their alignment, and version histories (Dumontier et al. 2017).

This paper presents work on creating a new kind of VIAF-like mapping system and semantic portal called SampoSampo for cross-domain CH collections. The novelty in our case is two-fold: The system is based on existing KGs and LOD services already available on the Semantic Web, and includes not only a LOD service but a semantic portal, too. The focus is on using data of the over 20 Sampo KGs (Hyvönen 2022) that publish LOD in different application domains of Cultural Heritage, such as artifacts (CultureSampo), literature (BookSampo), musical performances (OperaSampo), artworks (ArtSampo), parliamentary speeches and networks (ParliamentSampo), culturally significant people (BiographySampo), academic people registers (AcademySampo), military history (WarSampo), and epistolary collections (LetterSampo). These KGs constitute a component of the Finnish DARIAH-FI research infrastructure[6] (Hyvönen 2024). We present the underlying ideas of SampoSampo, methods used, and report first results of creating the SampoSampo in practise.

2. Use cases of SampoSampo

There are many reasons and use cases for creating SampoSampo. The entity resources used in the Sampo systems are often based on the same infrastructural resources available at the Finnish ontology services ONKI[7] and Finto[8]. However, due to historical and other reasons, also application specific KOS have been used for populating the metadata models, and the data is linked to international datasets, too. A goal of SampoSampo is to create a kind of universal reference service and a SPARQL endpoint for using the resources of Sampos. This kind of LOD service is useful when creating and aligning new datasets with existing ones by FAIR principles[9]. Furthermore, one benefit of collecting and assembling biographical data from multiple distinct sources is getting it enriched. As an example, one data source might have the main focus on someone’s career as a politician or as an artist while there might be more information about, e.g., her or his family relations in another data source. Besides that, comparing actor data imported from multiple databases also facilitates detecting contradictions and errors in the data sources.

We also aim at providing the end users with a single semantic portal and UI to a set of underlying KGs based on their shared resources. Using this portal, it is possible to search, browse, and analyse several local KGs on a global level at the same time. Furthermore, the portal will be capable for cross-cultural knowledge discovery of semantic association between entities (Lehmann, Schüppel, and Auer 2007; Tartari and Hogan 2018) and explaining them in natural language, using the “knowledge-based” approach to relational search (Hyvönen and Rantala 2021; Rantala, Hyvönen, and Leskinen 2023; Rantala et al. 2024).

Our focus is on resources for historical persons, organizations, and places that are widely used in virtually all Sampo systems. For example, Figure 1 illustrates the linkedness of some biographical Sampo systems and other systems, including the Kanto authority file system[10] by the National Library of Finland, Wikidata[11], and the German Integrated Authority File system of the Deutsche National Bibliothek (GND)[12]. The numbers on the connection arcs tell the number of shared resources between the connected datasets. For example, from the 100 000 persons in the WarSampo KG (Koho et al. 2021; Hyvönen et al. 2016) 2600 can be found in Wikidata and 290 in the ParliamentSampo KG (Leskinen, Hyvönen, and Tuominen 2021; Hyvönen et al. 2024).

Acknowledgements

This research was funded by the Research Council of Finland and is part of the FIN-CLARIAH initiative that has received funding from the European Union NextGenerationEU instrument. Computing resources provided by the CSC – IT Center for Science were used in our work.

PDF version of this abtract: https://seco.cs.aalto.fi/publications/2025/hyvonen-et-al-samposampo-abstract-2025.pdf

Endnotes

1 5-star model: https://5-star.info
2 Dublin Core metadata initiative: https://dublincore.org

[3] Resource Description Framework RDF: https://www.w3.org/RDF/
[4] Simple Knowledge Organization System SKOS: https://www.w3.org/2004/02/skos/
[5] Virtual International Authority File system: https://viaf.org
[6] LOD part of the FIN-CLARIAH/DARIAH infrastructure: https://seco.cs.aalto.fi/projects/fin-clariah/
[7] ONKI service: https://onki.fi
[8] Finto service: https://finto.fi
[9] FAIR principles: https://www.go-fair.org/
[10] Kanto authorities: https://finto.fi/finaf/en/
[11] WIkidata: https://wikidata.org
[12] Gemainsame Normdatei: https://www.dnb.de/EN/Professionell/Standardisierung/GND/gnd_node.html

Bibliography

Dumontier, Michel, Pierre-Yves Vandenbussche, Ghislain A. Atemezing, María Poveda-Villalón, and Bernard Vatant. 2017. "Linked Open Vocabularies (LOV): A gateway to reusable semantic vocabularies on the Web." Semant. Web (NLD) 8, no. 3 (January): 437–452. issn: 1570-0844. https://doi.org/10.3233/SW-160213. https://doi.org/10.3233/SW-160213.

Hickey, Thomas B., and Jenny A. Toves. 2014. "Managing Ambiguity In VIAF." DLib Magazine 20 (7/8). https://doi.org/doi:10.1045/july2014-hickey.

Hyvönen, Eero. 2022. "Digital Humanities on the Semantic Web: Sampo Model and Portal Series." Semantic Web 14 (4): 729–744. https://doi.org/10.3233/SW-223034.

Hyvönen, Eero. 2024. "How to Create a National Cross-domain Ontology and Linked Data Infrastructure and Use It on the Semantic Web." DOI: 10.3233/SW-243468, Semantic Web, https://doi.org/10.3233/SW-243468.

Hyvönen, Eero, Erkki Heino, Petri Leskinen, Esko Ikkala, Mikko Koho, Minna Tamper, Jouni Tuominen, and Eetu Mäkelä. 2016. "WarSampo Data Service and Semantic Portal for Publishing Linked Open Data about the Second World War History." In The Semantic Web – Latest Advances and New Domains (ESWC 2016), edited by Harald Sack, Eva Blomqvist, Mathieu d'Aquin, Chiara Ghidini, Simone Paolo Ponzetto, and Christoph Lange, 758–773. Springer–Verlag, May. https://doi.org/10.1007/978-3-319-34129-3_46.

Hyvönen, Eero, and Heikki Rantala. 2021. "Knowledge-based Relational Search in Cultural Heritage Linked Data." Digital Scholarship in the Humanities (DSH) 16 (Supplement_2): ii155–ii164. https://doi.org/10.1093/llc/fqab042. https://doi.org/10.1093/llc/fqab042.

Hyvönen, Eero, Laura Sinikallio, Petri Leskinen, Senka Drobac, Rafael Leal, Matti La Mela, Jouni Tuominen, Henna Poikkimäki, and Heikki Rantala. 2024. "Publishing and Using Parliamentary Linked Data on the Semantic Web: ParliamentSampo System for Parliament of Finland." In print, Semantic Web (October). https://seco.cs.aalto.fi/publications/2024/hyvonen-et-al-ps-swj-2024.pdf.

Koho, Mikko, Esko Ikkala, Petri Leskinen, Minna Tamper, Jouni Tuominen, and Eero Hyvönen. 2021. "WarSampo Knowledge Graph: Finland in the Second World War as Linked Open Data." Semantic Web 12, no. 2 (January): 265–278. https://doi.org/10.3233/SW-200392.

Lehmann, Jens, Jörg Schüppel, and Sören Auer. 2007. "Discovering Unknown Connections—the DBpedia Relationship Finder." In Proc. of the 1st Conference on Social Semantic Web (CSSW 2007), 113:99–110. LNI. GI. http://subs.emis.de/LNI/Proceedings/Proceedings113/gi-proc-113-010.pdf.

Leskinen, Petri. 2024. "Modeling and Using Biographical Linked Data for Prosopographical Data Analysis." PhD diss., Aalto University, School of Science, Department of Computer Science, October. https://seco.cs.aalto.fi/publications/2024/leskinen-phd-2024.pdf.

Leskinen, Petri, Eero Hyvönen, and Jouni Tuominen. 2021. "Members of Parliament in Finland Knowledge Graph and Its Linked Open Data Service." In Further with Knowledge Graphs. Proceedings of the 17th International Conference on Semantic Systems, 6-9 September 2021, Amsterdam, The Netherlands, 255–269. IOS Press. https://doi.org/10.3233/SSW210049.

Rantala, Heikki, Eero Hyvönen, and Petri Leskinen. 2023. "Finding and explaining relations in a biographical knowledge graph based on life events: Case BiographySampo." In Joint Proceedings of the ESWC 2023 Workshops and Tutorials co-located with 20th European Semantic Web Conference (ESWC 2023), vol. 3443. CEUR Workshop Proceedings. https://ceur-ws.org/Vol-3443/ESWC_2023_SEMMES_relations.pdf.

Rantala, Heikki, Petri Leskinen, Lilli Peura, and Eero Hyvönen. 2024. "Representing and searching associations in cultural heritage knowledge graphs using faceted search." In SEMANTiCS 2024, 20th International Conference on Semantic Systems, proceedings. In press. IOS Press, September. https://seco.cs.aalto.fi/publications/2024/rantala-et-al-searching-interesting-relations-2024.pdf.

Tartari, Gonzalo, and Aidan Hogan. 2018. "WiSP: Weighted Shortest Paths for RDF Graphs." In Proceedings of VOILA 2018, 37–52. CEUR Workshop Proceedings, vol. 2187. https://ceur-ws.org/Vol-2187/paper4.pdf.

Eero Hyvönen[1,2], Petri Leskinen[1], Henna Poikkimäki[1], Heikki Rantala[1], Jouni Tuominen[2,3], Senka Drobac[2], Ossi Koho[2], Ilona Pikkanen[4], Hanna-Leena Paloposki[4]

[1]Aalto University; [2]University of Helsinki (HELDIG); [3]University of Helsinki (HSSH); [4]Finnish Literature Society

Searching, exploring, and analyzing historical letters and the underlying networks: LetterSampo Finland (1809–1917) data service and semantic portal

Wednesday, 05/Mar/2025 2:30pm - 3:00pm

ID: 167 / Session LP 01: 3

Long paper (abstract) | 20-minute presentation with a 10-minute Q&A

Keywords: linked data, epistolary data, semantic portal, data analysis, network analysis

Abstract

Epistolary data (data related to letters) is by nature stored in geographically distributed archives and collections, as letters are exchanged between different people and places. To get a global view and analyze correspondences, the archival collections, and the underlying egocentric and social networks, data from the separate data silos in different cultural heritage (CH) organizations have to be aggregated, harmonized, and published as a global data service with APIs for Digital Humanities research and application development. This paper presents an overview of the system LetterSampo Finland (1800-1917) consisting of a Linked Open Data (LOD) service and a semantic portal designed for these purposes. The LOD service contains extensive metadata on one million letters sent or received in the Grand Duchy of Finland during 1809-1917, aggregated from various Finnish CH organizations, harmonized by using a shared ontological data model and vocabularies, and published as a LOO service with a SPARQL endpoint and data dumps under an open license. Based on the so-called Sampo model and Sampo-UI framework, a new semantic portal has been created on top of the LOO service. This portal can be used for searching, exploring, and analyzing letters, letter collections, and networks within these correspondences.

1. Introduction

Letters are an important source of data for historical research, biography, and prosopography. Letters have been in a central role for the development of scientific thinking: During the Age of Enlightenment it became possible for people to send and receive letters across Europe and beyond, based on a revolution in postal services. This opportunity resulted into the so-called

Republic of Letters (RofL) (Respublica litteraria), a cross-national collaborative communication network that formed a basis for modern European scientific thinking, values, and institutions in Early Modern times 1400-1800 (Hotson and Wallnig 2019; Miert 2016). Sending letters is a phenomenon that is in many ways analogous to many means of communication using the Internet, email, social media, and the World Wide Web (WWW) since the 1990' s (Urena-Carrion et al. 2022).

Collections of sent and received letters are therefore stored in various archives for future generations to study. To enable Digital Humanities (DH) research (McCarty 2005; Gardiner and Musto 2015) on heterogeneous, distributed letter collections, data about the letters have been aggregated, harmonized, and provided for the research community through various databases and web services. Examples of such services include Europeana[1], Kalliope[2], The Catalogus Epistularum Neerlandicarum[3], Electronic Enlightenment[4], ePistolarium[5], the Mapping the Republic of Letters project[6], SKILLNET[7], correspSearch[8], and the Early Modern Letters Online (EMLO) catalogue[9].

From a technical point of view, epistolary metadata are challenging as letters are distributed in different cultural heritage organizations, have been catalogued using different data models and vocabularies, the letters are written in different languages, and the collections are typically incomplete. Using linked data provides a promising approach to tackle these problems. In (Tuominen et al. 2022) application of the idea to the Early Modern Letter Online database of the Oxford University was discussed. The LetterSampo Framework for publishing and using epistolary linked data for DH research was introduced in (Hyvönen, Leskinen, and Tuominen 2023) and (Leskinen et al. 2024), and later employed in the Constellations of Correspondence - Large and Small Networks of Epistolary Exchange in the Grand Duchy of Finland (CoCo) project[10] for developing LETTERSAMPO FINLAND (1809-1917). This paper extends substantially our earlier paper on the CoCo project (Tuominen et al. 2022). We present results of applying and extending the LetterSampo Framework, using the so-called Sampo model (Hyvönen 2022) and Sampo-UI framework (Ikkala et al. 2022; Rantala et al. 2023), and the software developed for the case study regarding letters sent and received in Finland during the Grand Duchy in Finland (1809-1917).

2. Contributions of LetterSampo Finland (1809-1917)

1. Searching letters. If a researcher was looking for the letters written or received by a person X, it has not been easy to find out in what archives such letter can be found. For the first time, fundamental queries like "Find all letters sent (or received) by person X" can be answered under a single search engine in Finland.

2. Providing a global view of correspondences. The aggregated collection of LETTERSAMPO FINLAND (1809-1917) provides a global view of the scattered letters can now be provided. From a quantitative point of view it has not been possible to answer simple questions, such as "How many letters available in archives are were sent in Finland during a period X in 1809-1917".

3. Analyzing metadata. Based on the metadata, it is now possible to analyze correspondences in flexible ways based on, e.g., persons, times, and places. For example: "How many letters did X receive from Y during time T", "How many letters were sent to place X by person Y?", "Who are the most active letter writers?".

4. Analyzing letter content. In some well-curated collection of prominent Finns, such as Johan Vilhelm Snellman (1806-1881), Zacharias Topelius (1818-1898), Johan Ludvig Runeberg (1804-1877), and Elias Lönnrot (1802-1884), not only metadata but also the letter contents are available in digital form for for textual and other analyses.

5. Analyzing underlying networks. Sending and receiving letters indicate social networks underlying the correspondences. Network analysis can be used to find out, for example, egocentric networks of individual people, social networks of groups of people, their central figures (hubs), and to study processional and family communications, or how the networks evolve in time (temporal networks).

6 Developing infrastructure for epistolary data. The primary data and metadata in archives is available in various formats, such as PDF and Word documents, spreadsheets, and in different kind of databases. It would be important to develop shared data

models and vocabularies for representing epistolary data in the future based on the FAIR principles[11], so that the data would be more Findable, Accessible, Interoperable, and Re-usable in the future as the collections evolve and new ones are established.

7. Analysing archival collections The data can also be used for finding out what kind of epistolary fonds different archives have, how the collections have evolved in time, and to study geographical distributions of where the letters have been sent and received.

3. Paper outline

The full paper based on this abstract is organized as follows:
First, its is explained how a questionnaire was sent in Finland to over 100 CH organizations that were expected to host collections of letters from the time period c. 1800-1917. As a result, data from several organizations were received and statistics based on this investigation are provided in the paper.

Second, the tedious data cleaning process (Drobac et al. 2023) and pipelines for transforming primary dataset into linked open data are described. Several challenges were encountered: the data came in various heterogeneous forms that often needed human interpretation. Also issues of data quality, errors, and incomplete data arose. A major challenge here was linking and aligning person names with unique entities as person names change in time due to, e.g., marriages and deliberate name changes (Drobac, Leskinen, and Wahjoe 2023). Furthermore, various name variants have been used for the same persons in different archives and by different catalogers in different times. To tackle this, biographical data including, e.g., the times of living as well as the known name variations of individuals has been assembled from various data sources including earlier publications iin Sampo series. Furthermore, the person data is also enriched from these external sources.
Third, the ontology-based data model extended from that of (Leskinen et al. 2024) and well as the vocabularies used for populating Linked Data are overviewed. In the data model the classes in the most central role are the metadata records, the letter resources, and the actors in correspondences.

From a data perspective, a major challenge in the case study was that in many, if not most cases, letter-wise metadata were not available but only metadata about archival units. For example, a particular unit in an archive may contain N letters that two families exchanged during a time period T, but it is not known who sent what letter to whom. On the other hand, in some cases pertaining to people of national importance, very detailed metadata about individual letters, including content annotated with mark-up such as TEI[12] was available. Another challenge of the data is its size: the KG contains information about nearly one million letters coming from 11 archives and 1100 fonds, with 95 000 historical people and 2000 places referred to.

Fourth, the process of establishing the LOD service and SPARQL endpoint using the Linked Data Finland platform LDF.fi[13] and publishing the data dumps an part of the national FINCLARIAH research infrastructure[14] are explained.

The LOD service SPARQL API can be used directly for DH research by, e.g., the Yasgui SPARQL query editor (Rietveld and Hoekstra 2017) or Jupyter Notebooks[15]. Examples with visualizations of this use case are given. Furthermore, we present results of using network analysis on epistolary data, using, e.g., the egocentric network based on Elias Lönnrot's correspondences (Poikkimäki, Leskinen, and Hyvönen 2024).

Fifth, the LETTERSAMPO FINLAND (1809-1917) semantic portal with its four application perspectives based on integrating faceted semantic search and browsing with data-analytic tool is presented with examples for searching, exploring, and analyzing the underlying LETTERSAMPO FINLAND (1809-1917) knowledge graph.

Finally, contributions of the research are summarized, challenges of using data-analytic methods in analyzing incomplete epistolary data are discussed, and directions for further research are proposed.

Acknowledgments

Thanks to Johanna Enqvist and Matti La Mela for their earlier contributions to the CoCo project. This research was funded by the Research Council of Finland and is part of the FIN-CLARIAH initiative that has received funding from the European Union NextGenerationEU instrument. Computing resources provided by the CSC - IT Center for Science were used in our work.

Endnotes

[1] http://www.europeana.eu
[2] http://kalliope.staatsbibliothek-berlin.de
[3] http://picarta.pica.nl/DB=3.23/
[4] http://www.e-enlightenment.com
[5] http://ckcc.huygens.knaw.nl/epistolarium/
[6] http://republicofletters.stanford.edu
[7] https://skillnet.nl
[8] https://correspsearch.net
[9] http://emlo.bodleian.ox.ac.uk
[10] CoCo project homepage in Aalto University: https://seco.cs.aalto.fi/projects/coco/
[11] FAIR principles: https://www.go-fair.org/
[12] Text Encoding Initiative TEI: https://www.tei-c.org/
[13] Linked Data Finland platform: https://ldf.fi
[14] Linked datata part of FIN-CLARIAH/DARIAH-FI: https://seco.cs.aalto.fi/projects/fin-clariah/
[15] Jupyter Notebooks: https://jupyter.org/

Bibliography

Drobac, Senka, Johanna Enqvist, Petri Leskinen, Muhammad Faiz Wahjoe, Heikki Rantala, Mikko Koho, Ilona Pikkanen, et al. 2023. “The Laborious Cleaning: Acquiring and Transforming 19th-Century Epistolary Metadata.” In Digital Humanities in the Nordic and Baltic Countries. Publication, DHNB2023 Conference Proceeding, 5:248–262. 1. University of Oslo Library, Norway. https://doi.org/10.5617/dhnbpub.10669.

Drobac, Senka, Petri Leskinen, and Muhammad Faiz Wahjoe. 2023. “Navigating the Challenges of Deduplicating Actors in Historical Letter Exchanges.” In Proceedings of the 24th European Conference on Knowledge Management, 24:1694–1697. 2. Academic Conferences International Limited. https://doi.org/10.34190/eckm.24.2.1317.

Gardiner, Eileen, and Ronald G. Musto. 2015. The Digital Humanities: A Primer for Students and Scholars. Https://doi.org/10.1017/CBO9781139003865. New York, NY, USA: Cambridge University Press.

Hotson, Howard, and Thomas Wallnig, eds. 2019. Reassembling the Republic of Letters in the Digital Age. Göttingen University Press. https://doi.org/10.17875/gup2019-1146.

Hyvönen, Eero. 2022. “Digital Humanities on the Semantic Web: Sampo Model and Portal Series.” Accepted, Semantic Web, http://www.semantic-web-journal.net/content/digital-humanities-semantic-web-sampo-model-and-portal-series.

Hyvönen, Eero, Petri Leskinen, and Jouni Tuominen. 2023. “LetterSampo – Historical Letters on the Semantic Web: A Framework and Its Application to Publishing and Using Epistolary Data of the Republic of Letters.” Journal on Computing and Cultural Heritage 16 (1).

Ikkala, Esko, Eero Hyvönen, Heikki Rantala, and Mikko Koho. 2022. “Sampo-UI: A Full Stack JavaScript Framework for Developing Semantic Portal User Interfaces.” Semantic Web 13 (1): 69–84. https://doi.org/10.3233/SW-210428.

Leskinen, Petri, Javier Ureña-Carrion, Jouni Tuominen, Mikko Kivelä, and Eero Hyvönen. 2024. “Knowledge Graphs and Data Services for Studying Historical Epistolary Data in Network Science on the Semantic Web.” Under open review, Semantic Web, https://www.semantic- web- journal.net/content/knowledge- graphs- and- data- services- studying- historical-epistolary-data-network-science-1.

McCarty, Willard. 2005. Humanities Computing. Palgrave, London.

Miert, Dirk van. 2016. “What was the Republic of Letters? A brief introduction to a long history (1417–2008).” Groniek 204/205:269–287.

Poikkimäki, Henna, Petri Leskinen, and Eero Hyvönen. 2024. “Using Network Analysis for Studying Cultural Heritage Knowledge Graphs – Case Correspondence Networks in Grand Duchy of Finland 1809–1917.” Under review. August. https://seco.cs.aalto.fi/publications/2024/poikkimaki-et-al-coco-2024.pdf.

Rantala, Heikki, Annastiina Ahola, Esko Ikkala, and Eero Hyvönen. 2023. “How to create easily a data analytic semantic portal on top of a SPARQL endpoint: introducing the configurable Sampo-UI framework.” In VOILA! 2023 Visualization and Interaction for Ontologies, Linked Data and Knowledge Graphs 2023. CEUR Workshop Proceedings, Vol. 3508, October. https://ceur-ws.org/Vol-3508/paper3.pdf.

Rietveld, Laurens, and Rinke Hoekstra. 2017. “The YASGUI family of SPARQL clients.” Semantic Web – Interoperability, Usability, Applicability 8 (3): 373–383. https://doi.org/10.3233/SW-150197.

Tuominen, Jouni, Mikko Koho, Ilona Pikkanen, Senka Drobac, Johanna Enqvist, Eero Hyvönen, Matti La Mela, Petri Leskinen, Hanna-Leena Paloposki, and Heikki Rantala. 2022. “Constellations of Correspondence: a Linked Data Service and Portal for Studying Large and Small Networks of Epistolary Exchange in the Grand Duchy of Finland.” In 6th Digital Humanities in Nordic and Baltic Countries Conference. https://seco.cs.aalto.fi/publications/2022/tuominen-et-al-coco-dhnb-2022.pdf. CEUR Workshop Proceedings, March.

Ureña-Carrion, Javier, Petri Leskinen, Jouni Tuominen, Charles van den Heuvel, Eero Hyvönen, and Mikko Kivelä. 2022. “Communications Now and Then: Analyzing the Republic of Letters as a Communication Network.” Applied Network Science 7 (26).

Eero Hyvönen[1,2], Jouni Tuominen[2,3], Heikki Rantala[1], Petri Leskinen[1], Rafael Leal[1], Annastiina Ahola[1]
[1]Aalto University; [2]University of Helsinki (HELDIG); [3]University of Helsinki (HSSH)

How to create a Linked Open Data service and semantic portal for your own Cultural Heritage data

ID: 124
Half-day tutorial
Keywords: linked open data, semantic web, data services, semantric portals, data analysis

General motivation for the tutorial

To facilitate Digital Humanities (DH) research, following basic tasks need to be done: 1) Transforming, harmonizing, and aggregating Cultural Heritage (CH) data into a form suitable for research. 2) Publishing the data with APIs for research and applications. 3) Provision of tools and applications for DH analyses.

Semantic web technologies and Linked Open Data (LOD) provide a promising approach and tools for these tasks. A proof of concept of this is the "Sampo series" of over 20 Linked Open Data services of CH data with semantic portals on top them that are in use in Finland (Hyvönen, 2023). These systems are used by both DH researchers and the general public with up to over million end users in some Sampos. Semantic web technologies can in many cases be used with little experience in computer science but are still not so well known and utilized by DH researchers, although more and more LOD datasets are readily available on the Web. This tutorial is targeted to mitigate this challenge.

Learning objectives/outcomes

Since 2002, the Semantic Computing Research Group (SeCo) in the Aalto University and University of Helsinki has been involved in 1) developing a national semantic web infrastructure and 2) the Sampo series of Linked Open Data (LOD) services and semantic portals. This work goes on today as part of the national Finnish DARIAH-FI research infrastructure program, in relation to Aalto University's collaborations with DARIAH-EU. Based on lessons learned during this work, the goal of this tutorial is to explain in practice how the standards, models, tools, datasets, and portals developed can be re-used for creating new LOD services and applications for DH, based on one's own data available in different formats.

Coordinators

Researchers of the SeCo group: Eero Hyvönen (contact), Jouni Tuominen, Annastiina Ahola, Heikki Rantala, Petri Leskinen, Rafael Leal (cf. program draft below)

Format

Presentations and demonstrations. Advise for hands-on experiments.

Target audience

DH researchers with computation interests. No specific skills required.

More information

Following articles overview our research and lessons learned regarding 1) the national LOD infrastructure, 2) Sampo series of LOD services and semantic portals, and 3) the vision of using the Semantic Web for DH research towards knowledge discovery and Artificial Intelligence-based systems:

1. Eero Hyvönen: How to Create a National Cross-domain Ontology and Linked Data Infrastructure and Use It on the Semantic Web. *Semantic Web*, in press, 2024. DOI: 10.3233/SW-243468.
2. Eero Hyvönen: Digital Humanities on the Semantic Web: Sampo Model and Portal Series. *Semantic Web*, vol. 14, no. 4, pp. 729-744, 2023.
1. Eero Hyvönen: Using the Semantic Web in Digital Humanities: Shift from Data Publishing to Data-analysis and Serendipitous Knowledge Discovery. *Semantic Web*, vol. 11, no. 1, pp. 187-193, 2020.

1) "Sampo" is a mythical device playing a central role in the Finnish epic Kalevala. According to a popular interpretation, it is regarded as a metaphor of amazing ancient technology.

Bibliography

Eero Hyvönen: How to Create a National Cross-domain Ontology and Linked Data Infrastructure and Use It on the Semantic Web. Semantic Web, in press, 2024. DOI: 10.3233/SW-243468.

Eero Hyvönen: Digital Humanities on the Semantic Web: Sampo Model and Portal Series. Semantic Web, vol. 14, no. 4, pp. 729-744, 2023.

Eero Hyvönen: Using the Semantic Web in Digital Humanities: Shift from Data Publishing to Data-analysis and Serendipitous Knowledge Discovery. Semantic Web, vol. 11, no. 1, pp. 187-193, 2020.

Daniel Ihrmark, **Ahmad Kamal**
Linnaeus University, Sweden

Coding as Disciplinary Literacy for Digital Humanities

ID: 300 / **WS09: 3**
Workshop on Digital Humanities and Social Sciences/Cultural Heritage (DHSS/DHCH) in Higher Education
Keywords: programming, teaching, disciplinary literacy, computational thinking

Disciplinary literacy refers to the specific literacies required to understand and engage with materials written using the semiotic repertoire unique to a given field (Shanahan & Shanahan, 2017). Traditionally, it bridges language learning and content knowledge, emphasizing modes of communication, such as graphs and charts in the social sciences, that are integral to understanding specific classroom subjects. This proposal extends the concept of disciplinary literacy to include programming as an important literacy for Digital Humanities (DH) education, particularly when integrated with the established framework of computational thinking (Wing, 2006).

Over the past two years, cohorts in the Linnaeus University DH MA program have been introduced to Python scripts in Google Colab as a supplementary component to a first-semester course on digital methods. The course primarily relies on executables with graphical user interfaces (GUIs) for assignments in text analysis, network analysis, and Geographic Information Systems (GIS). Alongside these tools, Python scripts performing similar tasks at a more basic level have been provided within each module. In addition, a similar mode of teaching has been used during the BALADRIA summer schools, with a mixed group of PhD and MA students from different fields and disciplines.

The inclusion of Python scripts is not intended to primarily develop students' practical coding skills, but to foster engagement that builds towards literacy in reading and interpreting code in a similar way to how other introductions to languages are carried out. This foundational literacy aligns with Moje's 4Es of disciplinary literacy—engaging students in authentic disciplinary practices, eliciting and engineering prior knowledge and experiences, examining the underlying principles and methods, and evaluating their understanding to empower them as actors within the discipline (Moje, 2015). Integration with focus on reading and interpretation can also serve to concretize the decomposition, pattern recognition, abstraction, and algorithm steps suggested by computational thinking by showcasing how they result in working code (Wing, 2006).

By framing coding as disciplinary literacy, this presentation will argue for the deliberate integration of programming into DH education as a semiotic mode for engaging with content knowledge. The effects of such integration are discussed with a basis in student evaluations and instructors' teaching experiences.

Bibliography

Moje, E. B. (2015). Doing and teaching disciplinary literacy with adolescent learners: A social and cultural enterprise. Harvard Educational Review, 85(2), 254–278. https://doi.org/10.17763/0017-8055.85.2.254

Shanahan, C., & Shanahan, T. (2017). Disciplinary literacy. In Handbook of Writing Research (2nd ed.). Taylor & Francis.

Wing, J. M. (2006). Computational thinking. Communications of the ACM, 49(3), 33–35.
https://doi.org/10.1145/1118178.1118215

Anton Karl Ingason, **Lilja Björk Stefánsdóttir**
University of Iceland, Iceland

Extracting, annotating, and publishing Icelandic style-shift data

Thursday, 06/Mar/2025 2:00pm - 2:20pm
ID: 206 / Session SP 06: 1
Short paper (full-text) | 15-minute presentation with a 5-minute Q&A
Keywords: sociolinguistics, style-shift, parliament data, annotation, database publication

We describe the methods that we use as part of the ERC project Explaining Individual Lifespan Change in order to extract data from the Icelandic Parliament Corpus, manually annotate the extracted data for the purpose of analyzing sociolinguistic variables, and eventually publish the annotated database in open access under a Creative Commons Attribution License (CC BY 4.0) and using a well established repository that is suitable for the digital humanities (CLARIN). We furthermore discuss how the data that we extract from the parliament corpus gains more qualitative depth by us interviewing some of the politicians in question directly.

Gerth Jaanimäe
University of Tartu, Estonia

Improving the accuracy of normalizing historical Estonian texts by combining statistical machine translation with Bert language model

Friday, 07/Mar/2025 2:20pm - 2:40pm
ID: 177 / Session LP/SP 10: 3
Short paper (full-text) | 15-minute presentation with a 5-minute Q&A
Keywords: Natural language processing, machine learning, historical texts, normalization, non-standard spelling

Automatic analysis of historical texts is often hindered by different spelling system compared to the one used today. One approach to address this issue is to convert these texts to present-day spelling conventions, also called normalizing. One of the relatively old, however, still used methods for that is character level statistical machine translation (CSMT). As the method views the text one word at a time, it can lead to wrong normalizations due to lack of context. This paper gives an overview of integration of CSMT with BERT language models on texts written in Estonian, a morphologically rich language, to increase normalization accuracy.

Maciej Janicki[1], Helina Harend[2,4], Kati Kallio[1,3], Liina Saarlo[4], Mari Väina[4]
[1]University of Helsinki, Finland; [2]University of Tartu, Estonia; [3]Finnish Literature Society, Finland; [4]Estonian Literary Museum, Estonia

Computationally enhanced type indexing of Finnic oral poetry

Thursday, 06/Mar/2025 2:40pm - 3:00pm
ID: 225 / Session SP 07: 3
Short paper (abstract) | 15-minute presentation with a 5-minute Q&A
Keywords: folk poetry, oral tradition, text similarity, text reuse, type index

Several Finnic languages – North and South Estonian (including Seto), Votic, Ingrian (Izhorian), Karelian, Ludic and Finnish – share a common tradition of oral poetry called runosong, regilaul, Finnic alliterative tetrameter, Kalevalaic poetry, etc. This song tradition has been documented most extensively in the 19th-20th centuries resulting in large collections in Estonia, Finland and Russian Karelia. Finnish and Estonian databases contain c. 250 000 digitized poem texts relating to runosong tradition that have been brought together in the framework of the collaborative FILTER project since 2020.

The current presentation explores the possibilities of using a computational similarity detection method to complement the metadata of the song texts by type assignments in case they are missing.

The early 20th century research was highly interested in analyzing variation of individual poem types, and comparative studies. More recent research has mostly concentrated on regional or parish level traditions and individual singers. Although there have been some comparative efforts, there is, nevertheless, no overall understanding of how the different features, formulas, motifs and poem types are shared and differ across linguistic and cultural areas. The tradition is really versatile. While some short poem types may be relatively stable, the tradition was also characterized by a high degree of complex, multilevel variation. Many formulas and motifs can be used in several quite different poetic contexts, and the singers often merged different plots and motifs, creating their own versions.

Typological indexing of song texts has been inevitable in order to enable finding the materials with similar content and/or functions in large collections, and to get better insight into the nature of tradition. The principles of indexing, as well as type names have evolved throughout the history of folklore studies, and are not totally consistent. Most often, a poem type is identified by the common content or themes – or a plot in case of narrative poems – and some key motifs and line types, but in some cases rather by function or ritual context. Especially, the indexing of shorter fragments may be challenging if these consist of multifunctional verses and motifs. One text often consists of several motifs, and consequently is given several type names, whereas the correspondences of text parts with the type names are currently not indicated in the database.

The digitized text material in Finland and Estonia is divided into three historical corpora. The Estonian ERAB (Estonian Runosongs' Database) corpus of mostly Estonian poems and the Finnish SKVR corpus of Karelian, Ludic, Ingrian, Votic and Finnish poems (published as a book series 1908–1997) contain manually made poem type indices. The Finnish JR corpus of unpublished poems is parallel to SKVR corpus but does not have a type index. The present paper explores the ways to solve the task of indexing non-indexed parts of the corpora: part of ERAB and JR.

In a collaboration between folklorists and computer scientists, we have developed a method for automatically detecting similar texts in the collections. The method is based on bag-of-character bigrams vector representation of lines and line-wise weighted edit distance alignment of songs. The similarity computation together with the user interface Runoregi to present the results have been described extensively in our earlier publications [1,2,3].

In the present work, the text similarity detection method is applied to enhance the creation of a type index for yet unlabeled material. The method is based on the assumption that similarity in textual content is a sufficient (though not necessary) condition to assign the same type label(s). Initial explorations have shown that this process frequently leads to re-evaluating existing labelings as well. Thus, instead of using the labeled part of the collections as training data for a classifier, we decided for a more cautious approach, in which type labels are transferred between pairs of highly similar poems.

Our experiment includes two different scenarios for different collections. In the Estonian Runosongs' database, the type index is work in progress, with roughly 30% of the material being unlabeled. Also new texts are continuously added to the database. Here we use the similarity detection method to identify pairs of very similar poems, in which one is labeled and the other one is not. We assume that in such cases the labels can be transferred to the unlabeled poem. However, the result needs to be checked manually. To assess the usefulness of the method, we have selected a sample of pairs with different degrees of similarity, which was manually processed. The results show that the automatically generated labels are often correct. However, another typical case is that both poems in the pair are indeed labeled in the same way, but the classification of the already labeled poem needs to be changed.

On the other hand, the collection of unpublished poems (JR) of the Finnish Literature Society is completely unlabeled, but partly similar or overlapping with the SKVR collection which has a complete type index. We currently do not aim for obtaining a manually checked index for JR. Instead, we are interested in estimates on the collection's content. We thus identify pairs of a labeled SKVR poem and an (unlabeled) JR poem with high textual similarity and transfer the labels automatically. Further, we study how the estimate changes depending on the similarity threshold used. For JR poems with no SKVR counterpart we further aim to obtain an estimate by applying text clustering and manually labeling the largest clusters.

Labeling the unindexed parts of the material aids significantly both the close and distant reading of the material, as the type indices offer a quick way to search, understand and compare the main contents of the texts. The similarity detection method has a potential of significantly speeding up the manual annotation process.

Bibliography

[1] Maciej Janicki, Kati Kallio and Mari Sarv. Exploring Finnic written oral folk poetry through string similarity. In: Digital Scholarship in the Humanities 38:1, 2023.

[2] Maciej Janicki. Large-scale weighted sequence alignment for the study of intertextuality in Finnic oral folk poetry. In: Journal of Data Mining and Digital Humanities, 2023.

[3] Maciej Janicki, Kati Kallio, Mari Sarv and Eetu Mäkelä. Runoregi: A user interface for exploring text similarity in oral poetry. In: DHNB 2024.

Risto Järv
Estonian Literary Museum, Estonia

Analysing Fairy Tales with AI: Identifying Violent Deaths and Propp’s Functions

Wednesday, 05/Mar/2025 4:20pm - 4:40pm
ID: 268 / Session SP 02: 2
Short paper (abstract) | 15-minute presentation with a 5-minute Q&A
Keywords: AI, Fairy Tales, Propp’s Functions, Text Analysis

Fairy tales are narratives of overcoming obstacles and achieving success, often concluding with a happy ending. However, they also contain darker elements, including violence and cruelty. Instances of such violence appear as early as the Kinder- und Hausmärchen by the Brothers Grimm, where it served as a pedagogical tool. Similarly, violence is present in many of the 16,000 fairy tales archived at the Estonian Folklore Archives. This study explored the use of AI-based text analysis tools in detecting violent deaths in fairy tales, focusing on the Monumenta Estoniae antiquae series, which provides a typological overview of Estonian fairy tales. These volumes represent three distinct categories: wonder tales (Eesti muinasjutud; EMj I:1, I:2) with human protagonists, animal tales (EMj II) reflecting human behaviour through animal characters, and realistic tales (EMj III) featuring human protagonists. For the analysis, I selected 100 consecutive fairy tales from each volume.

When developing the analytical framework, I applied different AI-based chatbot tools available in mid-year 2024—ChatGPT, Microsoft Copilot, and Gemini Advanced—each with distinct processing capabilities. The instruction sets underwent multiple iterations, enhancing analytical precision. The final phase of analysis was completed using ChatGPT 4o. The AI was instructed to detect all instances of violence to verify whether any had been overlooked, and the models also estimated the proportion of violence in each text. I manually spot-checked the AI-generated results, primarily through event summaries and my own knowledge of fairy tales, to refine the identification of violent incidents. The results showed that analysing each tale as a complete narrative yielded the most accurate outcomes. AI comprehension improved when texts were first segmented into sentences, allowing for more precise detection of violent events.

Across the 300 analysed fairy tales, 268 violent death events were recorded: 133 in wonder tales, 79 in animal tales, and 56 in realistic tales. Violent deaths occurred in 70% of wonder tales and over 40% of animal and realistic tales. Animal tales emerged as the most violent subgenre, as all deaths in this category were violent, whereas wonder tales sometimes depicted natural deaths. The most frequently assigned Proppian function was PF 30 ("Punishment"), accounting for one-third of violent deaths in wonder tales, primarily involving the retribution of antagonists or false heroes. Other common functions included PF 8 ("Villainy") and PF 18 ("Victory"). In animal tales, PF 8 was linked to over half of all violent deaths, indicating that these tales focus less on moral resolution than wonder tales. The findings suggest that while AI can effectively process large datasets, with the current stage of analysis, human oversight remains essential to ensure accuracy.

Despite facilitating rapid data processing, AI did not always meet expectations on the first pass. Initial errors included misclassifying events, failing to detect key sentences, and overassigning Propp’s character roles—frequently attributing functions to background information or summaries. Manual corrections were necessary to refine the dataset. Although the AI’s ability to classify fairy tale elements improved with refined instructions, manual validation and corrections were often required, underscoring the limitations of automated literary analysis. The study highlights both the potential and the challenges of AI-assisted corpus analysis in folklore research. While AI-driven models offer valuable insights into large text collections, they still require human interpretation to refine results, particularly when dealing with nuanced or metaphorical descriptions of violence.

The findings suggest that while AI can effectively process large datasets, with the current stage of analysis, human oversight remains essential to ensure accuracy. Although the AI's ability to classify fairy tale elements improved with refined instructions, manual validation and corrections were often required, underscoring the limitations of automated literary analysis.

Bibliography

Hatzel, H. O., Stiemer, H., Biemann, C., & Gius, E. (2023). Machine learning in computational literary studies. Information Technology, 65(4–5), 200–217.

Olrik, A. (1992). Principles for oral narrative research. K. Wolf & J. Jensen (Trans.). Indiana University Press.

Propp, V. (1968). Morphology of the folktale. L. Scott (Trans.), L. A. Wagner (Ed.). University of Texas Press.

Ziems, C., Held, W., Shaikh, O., Chen, J., Zhang, Z., & Yang, D. (2024). Can large language models transform computational social science? Association for Computational Linguistics.

Atli Jasonarson, **Steinþór Steingrímsson**
The Árni Magnússon Institute for Icelandic Studies, Iceland

I Can Explain This: Enhancing Grammatical Error Correction with Explanatory Feedback in Icelandic

Wednesday, 05/Mar/2025 5:00pm - 5:20pm
ID: 161 / Session SP 04: 4
Short paper (full-text) | 15-minute presentation with a 5-minute Q&A
Keywords: grammatical error correction, computer assisted language learning, automated written corrective feedback, Icelandic

We present an automated written corrective feedback system which connects a grammatical error correction model and the Icelandic Official Spelling Rules. The system is intended to aid users in correcting their spelling as well as provide them with explanations for the suggestions the model makes, alongside references to the relevant rules in the spelling rules. By providing the users with these references, we hope to help them to gain a better understanding of the Icelandic spelling rules. We compare our system to two large language models, GPT-4o and Claude 3.5 Sonnet, and discuss how it compares to them when it comes to correcting an erroneous text as well as explaining their corrections. Furthermore, we demonstrate that these large language models exhibit a problematic tendency to offer incomplete or even inaccurate explanations for their edits.

Ellert Johannsson
The Árni Magnússon Institute for Icelandic Studies, Iceland

Leveraging ChatGPT in analyzing legacy dictionary data

ID: 234
Short paper (abstract) | 15-minute presentation with a 5-minute Q&A
Keywords: ChatGPT, Legacy Dictionaries, Lexicography, Semantic Analysis

The integration of advanced AI tools like ChatGPT presents innovative opportunities for understanding and enhancing the data found in legacy dictionaries. This presentation explores the application of ChatGPT in the context of analysing structured data from older dictionaries, focusing on their value as representatives of the language and culture at the time they were compiled, as well as their semantic content and enhancement through additional target languages.

A case study for this approach is Björn Halldórsson's trilingual dictionary, which emerged during a significant shift in Icelandic lexicography. While the 17th and 18th centuries saw efforts primarily focused on Old Norse/Icelandic and Latin, Halldórsson's dictionary, published in 1814, marks a transition to modern languages, by including Danish as a target language. With around 29,000 entries, this dictionary provided both Latin and Danish translations of Icelandic headwords. The dictionary was republished in 1992 and as a result also exists in a digital format (as TeX-documents).

Halldórsson utilized a range of sources, including medieval texts and earlier dictionaries, but also drew from his own contemporary Icelandic language, incorporating vocabulary related to agriculture, trades, and daily life. As a result the headwords of the dictionary represent a rather broad semantic range that ChatGPT can further quantify by analyzing the dictionary's content and assigning words and senses to a system of semantic domains.

The presentation will explore some of the ways that ChatGPT can be useful in analysing the data from Halldórsson's dictionary. By employing natural language processing (NLP) techniques, ChatGPT can extract definitions, example sentences, and related terms from the digtal version of Halldórsson's dictionary. It can identify gaps or inconsistencies within the data. Furthermore, ChatGPT can facilitate the identification of new lexical information, drawing on contemporary language data to illustrate how words have developed or changed in meaning over time. Some examples of this will be shown in the presentation.

An additional transformative capability of ChatGPT lies in its machine translation functionalities. In the context of Halldórsson's dictionary, for instance, ChatGPT can introduce English as an additional target languages, generating equivalent definitions and examples that broaden the dictionary's utility for potential users. It can also compare the definitions given in Latin vs. Danish and identify inconsistencies and discrepancies, as is evident from some of the examples discussed.

The use of ChatGPT in analysing lexicographic data offers an innovative approach to disseminating and understanding the material found in legacy dictionaries like Halldórsson's. By drawing on this LLM's capabilities to analyze structured data, incorporate semantic domains, and add multiple target languages the user may gain a deeper insight into the linguistic and cultural material presented in such works while also noticing novel aspects of later language developments. It is important to keep in mind that while ChatGPT can generate suggestions and highlight potential areas of interest, human oversight is necessary to ensure accuracy and relevant context.

The study demonstrates how ChatGPT can be used in various manners to analyse legacy lexicographic data and through the processes described above provide a deeper understanding of the contemporary language of the time of the dictionary compiling as well as shedding some light on later languge development.

Bibliography

Andrésson, Guðmundur. 1999. Lexicon Islandicum. Orðabók Guðmundar Andréssonar. [New edition]. Gunnlaugur Ingólfsson & Jakob Benediktsson [eds]. Orðfræðirit fyrri alda IV. Reykjavík: Orðabók Háskólans.

Árnason, Jón. 1738. Nucleus latinitatis, Quo; Romani sermonis Vocis, ex classicis Auctoribus aurea argentineæ; atatis, ordine Etymologico adducta, et Interpredatione vemacula expositæ comprehenduntur. In usum Scholœ Schalholtinœ.

Árnason, Jón. 1994. Nucleus latinitatis [...]. [New edition]. Guðrún Kvaran og Friðrik Magnússon (eds). Orðfræðirit fyrri alda 3. Orðabók Háskólans, Reykjavík.

Halldórsson, Björn. 1992. Orðabók. Íslensk-latnesk-dönsk. [New edition]. Jón Aðalsteinn Jónsson (ed.). Orðfræðirit fyrri alda 2. Reykjavík: Orðabók Háskólans.

Halldórsson, Björn. 1814. Lexicon Islandico-Latino-Danicum. Rasmus K. Rask (ed.). Havniæ: J. H. Schubothum.

Ingólfsson, Gunnlaugur. 2012. "De første Trykte Islandske ordbøger". in Birgit Eaker, Lennart Larsson, Anki Mattisson (eds.) Nordiska Studier i Leksikografi 11: Rapport från Konferensen om lexikografi i Norden Lund 24-27 maj 2011, 309-318.

Ólafsson, Magnús of Laufás. 2010. Specimen lexici runici and Glossarium priscæ linguæ danicæ [originally published in 1650]. Anthony Faulkes and Gunnlaugur Ingólfsson [eds]. Orðfræðirit fyrri alda V. Reykjavík: Stofnun Árna Magnússonar í íslenskum fræðum & London: Viking society for northern research, University College.

Lars G Bagøien Johnsen[1], **Jennifer Thøgersen**[2], **Live Rasmussen**[2]
[1]National Library of Norway, Norway; [2]VID Specialized University

Structuring OCR using LLMs for catalogue cards

ID: 183 / Poster Session 2: 15
Poster and demo (abstract) with accompanying a 1-minute lightning talk
Keywords: ocr, ai, catalogues, structure, llm

In this poster, we present the preliminary phase of a project to digitize the card catalog for the Lars Dahle Library at VID Specialized University. The goal of the overall project is to convey the story of Dahle, a missionary and scholar who was an active voice in the Norwegian religious community in the late 19th and early 20th centuries. While traditional library digitization tools can handle basic metadata extraction, they often struggle with historical catalog cards that contain complex annotations, non-standard abbreviations, and multiple languages. Our approach combines Large Language Models (LLMs) with structured prompts to handle these challenges while maintaining consistency across the collection of over 4,000 books.

The workflow consists of three main stages. First, the scanned cards are converted from PDF to PNG format and then processed through pytesseract, a Python version of Google's Tesseract OCR engine, to extract text. The output of this stage is a CSV file with two columns: the card filename and the extracted text.

In the second stage, we use carefully crafted prompts to direct LLMs (Llama and Claude) in extracting and normalizing specific data points. These prompts were developed through interactive sessions with Claude, allowing us to iteratively refine the instructions based on real-time feedback and test cases before finalizing them for batch processing. The prompts guide the LLMs in:

1. Signature Codes: Through structured prompts, the LLM identifies classification markers like "ST," "Le," or "Te" from a predefined list of valid codes.

2. Author and Title Identification: The LLM processes OCR output using pattern-matching prompts, successfully handling cases with annotation standards (e.g., extracting "Carl Scharling" from "S(charling), C(arl)") and OCR errors.

3. Data Normalization: Using a combination of rule-based prompts and LLM capabilities, we:

- Standardize author names based on defined formatting rules
- Expand place name abbreviations (e.g., "Kbh." to "København/Copenhagen")
- Add geolocation data through a semi-automated verification process

The resulting structured data is organized into a knowledge graph, where entities (authors, publications, places) and their relationships are explicitly modeled.

The final stage involves human validation of the LLM outputs, particularly for ambiguous cases and geolocation data. The resulting structured data enables both traditional catalog searching and novel visualization approaches, which we demonstrate using Python's folium package to create interactive maps of publication locations.

This hybrid approach of combining LLMs with structured prompts and human oversight offers advantages over traditional catalog digitization tools, particularly in handling historical materials with non-standard formatting and multiple languages. The knowledge graph representation is particularly valuable for capturing the rich interconnections in historical library collections, such as tracking publication locations over time or identifying networks of authors and publishers. In addition, the techniques developed here can be extended to other library digitization projects, especially those involving documents with rich metadata like book colophons.

Bibliography

Brown, T., et al. (2020). Language Models are Few-Shot Learners. arXiv preprint arXiv:2005.14165.

Touvron, H., et al. (2023). LLaMA: Open and Efficient Foundation Language Models. arXiv preprint arXiv:2302.13971.

Smith, R. (2007). An Overview of the Tesseract OCR Engine. In Proceedings of the Ninth International Conference on Document Analysis and Recognition (Vol. 2, pp. 629-633). IEEE.

Ilkka Jokipii, Juho Laulajainen
National Archives of Finland, Finland

Echoes of War: AI-Driven Insights from Finnish Jahvetin kirjelaatikko Archive

ID: 217
Poster and demo (abstract) with accompanying a 1-minute lightning talk
Keywords: HTR, AI, History, Archives, NER, Topic Modeling

The Labor Organisation of Brothers-in-Arms initiated activities aimed at maintaining and developing, among other things, spiritual national defense work and combating organized rumor campaigns. The activities expanded to include monitoring the effects of citizens' moods and political events. For this purpose, a correspondence service was established, which, after the start of the Continuation War, expanded to include letters being featured in radio broadcasts. Yrjö Kilpeläinen was chosen as the editor of these broadcasts. The program was very popular, and around 100,000 letters were sent during the war years. About 30,000 of these letters, deemed to be of broader interest, were addressed in the radio broadcasts, while responses to others were sent by mail. The letters covered a wide range of citizens' problems during the difficult conditions of the war and shortage periods, with the help of experts from various fields used in drafting the responses.

At the end of 1941, the correspondence service was transferred directly under the State Information Agency. The program, named "Jahvetin kirjelaatikko" ("Jahvetti's Mailbox"), was broadcasted from 1941 to 1945. The State Information Agency was dissolved at the end of 1944, and a State Information Center was established under the Prime Minister's Office, along with an Information Supervision Department within the Ministry of the Interior, to handle the reduced tasks. The Jahvetin kirjelaatikko archive was transferred to the National Archives in December 1958.

Our poster describes various computational methods for effective processing of large, digitized archives. The objective is two-fold. First, we introduce Jahvetin kirjelaatikko archive, a digitized historical collection of around 100 000 civilian sent letters. Second, we present various computational tools and methods particularly appealing in cases where there is a substantial mass of data to make sense of.

We explain the workflow from transforming raw digitized data into machine-readable format with Handwritten Text Recognition (HTR) such as the NAF Multicentury HTR pipeline. The pipeline is built on top of three machine learning models trained primarily on handwritten Finnish and Swedish texts from the 17th to 20th centuries, processing documents with a Character Error Rate (CER) of 3,1%. Grounded by practical examples, we further discuss how easy-to-employ yet powerful computational tools can streamline and strengthen research processes. Specifically, we explore how 1) Topic Modeling can label documents based on their content, providing both granular view of individual documents and a broader understanding of the collection, 2) Named Entity Recognition (NER) can enrich document metadata by extracting additional information like names, locations, dates, and events, and 3) Geotagging documents can reveal regional variations, providing insights into how events unfolded or impacted different areas.

By combining Topic Modeling, Named Entity Recognition, and Geotagging, we highlight that researchers can organize vast historical collections into manageable units and identify hidden patterns or discourses using different computational tools. They provide a supplementary yet effective way to analyze digitized archives, such as Jahvetin kirjelaatikko, enabling researchers to draw novel conclusions of the past phenomena yet to be discovered.

Sander Jürisson[1], **Anna-Liisa Õispuu**[2], **Teele Siig**[2]
[1]Freelancer, Estonia; [2]Estonian Maritime Museum

From Exhibit to Experience: Enhancing Learning with Museum Collections through Open Innovation

ID: 295 / **WS02: 7**
Innovations and New Interactions through Digital Cultural Heritage in GLAM Sector
Keywords: Participatory business models in the GLAM sector; Digital Storytelling and AR; Open innovation in the GLAM sector, Living Lab methodology

Museums are increasingly utilising their collections to reach audiences digitally, enhancing educational experiences for students and visitors alike. The Estonian Maritime Museum (EMM) has adopted this approach through two recent initiatives that expand digital learning and engagement. As a partner in the Horizon Europe RECHARGE project, EMM uses a Living Labs and open innovation framework to co-create classroom resources with partners from educational technology community. This collaborative process has enabled EMM to bring museum content to life in educational settings, offering students immersive, curriculum-aligned materials that connect them to maritime history interactively.

In another open innovation initiative, EMM served as a piloting partner with Ericsson, helping to design the user-focused aspects of the *Every Place Has A Story* platform. This augmented reality (AR) solution aimed to add digital storytelling layers to the museum's physical exhibits, enhancing visitor engagement by connecting artefacts to vivid historical narratives. Although the project ultimately proved unviable for long-term use, it offered EMM invaluable experience in digital storytelling based on museum collections, building the team's capacity to engage in similar future projects.

This presentation will explore EMM's experience with RECHARGE, illustrating how Living Labs support sustainable educational tools by merging community insights with technical expertise. It will also reflect on the lessons learned from the Ericsson collaboration, demonstrating how digital layers can transform museum collections both in-person and remotely, fostering impactful connections between cultural heritage and technology for diverse audiences. Additionally, the presentation will provide practical guidance on using the participatory cultural business model canvas developed during the RECHARGE project and offer insights into building a development process based on the Living Lab methodology.

Mustafa Kamal
Linnaeus University, Sweden

Atheist Videos and Islam: An Analysis of Ten Pakistani Atheist Videos

ID: 113 / S01: 8
Doctoral Consortium
Keywords: Atheism, Pakistan, Islam, Digital Media

Similar to other Muslim-majority contexts (Duile 2017; Elsässer 2021; Schäfer 2016; Schielke; Van Nieuwkerk 2019), atheism in Pakistan is gaining visibility, with digital platforms playing a central role in facilitating critiques of Islam by Pakistani atheists. My PhD thesis examines *'Digitalization and its Role in Atheist Activism in Muslim-majority Contexts: A Case Study of Pakistan.'* The significance and originality of this study stem from both its focus on a relatively underexplored topic—online atheist activism in Pakistan as an important part of the Muslim-majority contexts—and its methodology, which employs large-scale, computer-assisted analysis of naturalistic data.

Following a presentation of my thesis at DASH Introductory Workshop inUmeå University in September 2024, I now plan to present a section of my thesis at DHNB 2025. This presentation, which will constitute as the third chapter of my thesis, will offer a detailed thematic analysis of anti-religious/Islamic discourse in 10 videos produced by atheist activists, which were purposively sampled for an initial investigation to identify the themes and topics.

The aim of identifying and analyzing these themes is to contextualize them within broader Islamic Studies scholarship. I seek critical feedback on the methodology and the thematic areas under discussion, to refine both the analysis and theoretical framework.

Bibliography

Duile, T. (2017). Indonesian secularists and atheists live under the shadow of stigma. The Conversation. Retrieved from https://theconversation.com/indonesian-secularists-and-atheists-live-under-the-shadow-of-stigma-82697

Elsässer, S. (2021). "Arab Non-believers and Freethinkers on YouTube: Re-Negotiating Intellectual and Social Boundaries." *Religions*, 12(2).

Schäfer, S. (2016). Forming 'Forbidden' Identities Online: Atheism in Indonesia. ASEAS - Austrian Journal of South-East Asian Studies, 9(2).

Schielke, S. (2015). Egypt in the Future Tense: Hope, frustration, and ambivalence before and after 2011. Bloomington: Indiana University Press.

Van Nieuwkerk, K. (2019). Understanding Unbelief in Egypt: Report on Preliminary Findings. Radboud University, Nijmegen, the Netherlands.

Antti Kanner, Veronika Laippala
University of Turku, Finland

Exploring modal syntax across the written–spoken language continuum

Thursday, 06/Mar/2025 3:00pm - 3:20pm
ID: 216 / Session SP 07: 4
Short paper (abstract) | 15-minute presentation with a 5-minute Q&A
Keywords: Syntax, corpus linguistics, online discussions, computational methods

From a linguistic perspective, online discussion forums are spaces where language has many opportunities for relatively unrestricted variation.There, language users are free to alter the mode of their expression according to the needs arising in the evolving context and communicative environment. Further, online discussion forums seldom have very specific guidelines concerning how ardently the forum users should adhere to the norms of standardized language. A common consequence of this is that online discussion forums fluctuate relatively freely between colloquial, speech-like written language and more careful, standardized expression (Crystal 2011). As language users seek to emulate spoken language in writing, the language use begins to carry features related to idiolectal, gender, generational and geographical variation – basically along all the axes often used to characterise language use in sociolinguistics.

In this paper, we will study how syntactic features behave in three different Finnish language corpora (a news corpus, a spoken language dialect corpus and an online discussion forum corpus) and whether their distributions align against the expected axes of variation, showing first and foremost the distinction between standardized written language and more colloquial modes of expression in addition to common sociolinguistic background variables. Large-scale, distributional approach – building characterisations of linguistic expressions use with the help of features of their aggregated occurrences – has generally been relatively rare in studies on syntax (Dunn 2024). Whatmore, syntactic variation is probably the least studied aspect of linguistic variation, possibly because it requires quite large datasets and, regarding syntactic features, it is not easy to identify the competing variants the language users make their choices from. In corpus linguistic studies on register - situationally defined texts with specific aims, such as news articles (Biber 1988) - syntactic features have, however, been frequently used. This is because the level of syntax is intimately connected to the communicative aims associated with a register, - these features first and foremost seen as a vehicle of the communicative aims.

The questions we seek to answer are 1) can syntactic features distinguish between spoken language and written language corpora, 2) can sociolinguistic variables be used to explain variation in dialect and online discussion forum corpora, and 3) do syntactic constructions have identifiable patterns of use in each corpora? In order to make the study of syntactic variation more tangible, we restrict our focus to Finnish modal syntax (Kangasniemi 1992). We analyse modal verbs and the construction types formed around them, as well as modal adverbs. While modal expressions have some nuanced semantic differences between them, their fundamental modal functions make them comparable and thus possible to project as a range of variants where language users make their choices from.

The datasets used will be 1) a collection of news articles from four Finnish news outlets (broadsheet Helsingin Sanomat, tabloid Iltalehti, public broadcast company Yle and news agency STT), 2) the dialect corpus of the Finnish Syntax Archives, a spoken-language corpus collected for sociolinguistic purposes and manually annotated and transcribed, and 3) a corpus consisting of Finnish online discussion forum Suomi24. With these three corpora we aim to triangulate between carefully edited standardized news discourse (news articles), spoken language solicited in a interview-like setting (the dialect corpus) and free-form, unrestricted online discourse (Suomi24). We are especially interested in how the online forums blend features from literate and spoken features in their use of modal syntax: how, where and for what purposes either standardized or spoken forms are used and whether the variation can be connected to sociolinguistic variables.

Methodologically we will approach these questions by (1) using Mutual Information to assess mutual predictability between the studied feature sets that represent the use of modal syntax in the datasets and by (2) looking for correlations between speaker background variables and their use of modal features. Especially we are interested in background variables related to location, as it would connect the modal syntax to dialectal variation. The third question is approached by building a multinomial regression model of a sample of occurrences of modal features in each dataset and assessing how distinct their uses are in each of them. Such an approach has an established tradition in corpus linguistics and has been applied to, for example, Finnish mental verbs by Arppe (2008).

Bibliography

Arppe, A. (2008). Univariate, bivariate, and multivariate methods in corpus-based lexicography: a study of synonymy. University of Helsinki.

Biber, D. (1988) Variation across speech and writing. Cambridge University Press.

Crystal, D. (2011) Internet linguistics: a student guide. Routledge.

Dunn, J. (2024). Computational construction grammar: A usage-based approach. Cambridge University Press.

Kangasniemi, H. (1992) Modal Expressions in Finnish. SKS.

Andres Karjus
Tallinn University; Estonian Business School

AI-powered analytics for communication management and media monitoring

ID: 194 / Poster Session 2: 16
Poster and demo (abstract) with accompanying a 1-minute lightning talk
Keywords: artificial intelligence, large language models, communication, corpora, media data

The increasing capabilities of large language models and multimodal vision-language models (LLMs, VLMs) enable scaling of textual and visual data analytics, including that informing communication management and planning. Empirical data from media, social media, or surveys allows informed decision making, but manual analysis has been a bottleneck. Past supervised machine learning approaches required large training sets to be annotated, while communication work often involves sets of several complex societal and cultural variables that may be unique to each new question. Pretrained generative models enable automating qualitative tasks via the zero-shot approach where an instructable assistant model is tasked with analysis or annotation tasks on the fly. This contribution advocates using modular feature-analytic approaches over traditional holistic methods, and emphasizes the need for theory-driven questions, systematic unitization, and rigorous statistical modeling of the outcomes before drawing final qualitative expert conclusions.

Several concrete real-world case studies, accompanied by evaluation metrics, are presented to illustrate how LLMs can assist in tasks like communication and reputation monitoring, stance detection, media narrative quantification, visual trend analytics and survey interpretation, based on recent academia-industry intersectoral collaborations. These applied, practically oriented results complement the growing literature in academic computational social science and digital humanities that has demonstrated how even off the shelf LLMs are already quite capable of analyzing data in multiple languages, modalities and socio-culturally complex issues.

These findings raise implications for the teaching and practice of communication management: traditional qualitative analytics can now be largely automated, while making use of big data results (and applying models, if not provided as a bespoke application or tool) requires competencies not commonly taught in communication curricula. Limitations and challenges of this approach and current LLM technology will also be discussed. While LLMs provide scalability, we emphasize the need for human expertise for meaningful interpretation and translating analysis into actionable insights.

Jaagup Kippar
Tallinn University, Estonia

Finding more relevant texts using keywords

ID: 319 / WS04B: 1
Explorations of the dynamics of cultural phenomena in text corpora
Keywords: podcast, keywords

Keyword algorithms ie Simple Math or Text Rank give keyness scores to words in texts. Usually a bigger score means also bigger importance in text. Texts top keywords list sorted by frequency also gives a good overview of text corpus. But finding texts by keyword by score or by rank sometimes does not give enough relevant texts to this keyword. Hands-on workshop will share experience about combined searching from cultural podcasts (44000 automatically transcribed texts).

Kamilla Kärrstedt, **Rico Simke**
Institute for Language and Folklore, Sweden

Advancing Citizen Science and Digital Accessibility: The Development of Folke, Isof’s Digital Archive Platform

ID: 311 / WS10: 3
Tradition Archives Meet Digital Humanities II
Keywords: Digitalization, Citizen science, Transcription methods, Archive accessibility, User-centered development

The Institute for Language and Folklore (Isof) launched the digital archive platform *Folke* in January 2022, making digitized archival records, predominantly folklore records from the beginning of the 20th century, accessible to all. Since then, thousands of digitized records have been added to the platform, which now includes dialect texts and dialect audio recordings as well.

Since the beginning, crowdsourcing and citizen science have been fundamental components in the Folke endeavor. In brief, Folke is technically designed to be both a service for the public and for researchers, for searching, browsing and downloading records – i.e. a digital archive of sorts – as well as a platform for contributing to the collections via transcribing archive records or adding information about collectors, interviewees, locations etc.

One of the project group’s primary tasks since Folke’s launch has been to develop and improve transcription methods. Enthusiastic volunteers – citizen scientists – have transcribed over 34,000 pages since 2022. As a result, the material is now more searchable than ever before, and it is also possible to have it read aloud, machine-translated, automatically annotated, and more. Furthermore, the volunteers have helped activate and disseminate the collections; people who transcribe records on Folke use the texts in various ways – in articles, lectures, podcasts, for teaching, and on social media.

Due to the increased activity surrounding the archive records, we have encountered various questions along the way. Central aspects of our workflow include determining which records we can and should make accessible, and how best to present them to the users. This involves adapting Folke both to meet the needs of the users and to align with the nature of the material itself.

Citizen science has shown immense potential for Isof. We now aim to further expand the opportunities for the public to contribute to the accessibility of archive collections, thereby streamlining the work involved. New transcription methods are the main focus, and they revolve around two key paths:

- Updated and improved method for transcribing text material.

- Content registration and/or transcribing dialect audio recordings.

Isof’s registers contain information about where audio recordings were made and who made them. However, information about the content is often incomplete, or even missing entirely in many cases. The data about interviewees could also benefit from being more detailed.

Contextualizing the collections, which becomes increasingly important as more and more archive records become digitally available on Folke, is and will remain an ongoing activity. This, as well as the previously described activities, is accompanied by continued efforts to make Folke a sustainable infrastructure for the archive records as well as for the user experience and involvement.

During the workshop at DHNB, we will delve deeper into these subprojects, sharing insights and practical examples. Given our user-centered development approach, we also plan to highlight the perspectives of the citizen scientists and discuss how the Folke team continuously communicates with the contributors, building trust and promoting further commitment.

Olof Karsvall[1], Karl-Magnus Johansson[1], Dick Kasperowski[2]
[1]Swedish National Archives, Sweden; [2]University of Gothenburg

AI in Archival Research: The Need for Enhanced Digital Infrastructure in Digital Archives

Wednesday, 05/Mar/2025 5:00pm - 5:20pm
ID: 262 / Session SP 01: 4
Short paper (abstract) | 15-minute presentation with a 5-minute Q&A
Keywords: Digital Archives, Handwritten Text Recognition, Citizen Science

One of the core missions of archival institutions is to make archival information accessible for research and public use. In recent decades, this mission has increasingly focused on digital availability through searchable metadata and the digitization of analog documents. The Swedish National Archives exemplifies this effort, having developed a National Archive Database that includes metadata for 250,000 archives and provides access to over 200 million scanned analog documents. However, a significant limitation remains: users cannot search within the content of these scanned documents, and the metadata is often limited to general descriptions of archival collections, lacking details about individual volumes or documents.

Recent advancements in artificial intelligence (AI) are now revolutionizing archival and collections-based research. In Sweden, the National Archives is leading an ambitious initiative that harnesses machine learning to extract information from millions of scanned handwritten and printed documents. This shift has far-reaching implications, transforming the relationships and dynamics between technology, researchers, archivists, and the public. It not only reshapes how archives are accessed and utilized but also calls for an evolution in the digital infrastructure of archival institutions. Systems must now accommodate machine-generated data while enabling researchers and the public to contribute by co-developing AI models and improving AI outputs.

This paper synthesizes three years of experience integrating Handwritten Text Recognition (HTR) technology and citizen science approaches within the Swedish National Archives. A key objective has been to train a base HTR model for historical Swedish texts, engaging volunteers, such as genealogists, and experts in language and manuscript interpretation. Additionally, we have established a collaborative network called the “Transcription Node”, where researchers and GLAM institutions (Galleries, Libraries, Archives, Museums) contribute to the collective training of a large-scale HTR model named “The Swedish Lion”. This model is publicly available on platforms such as Transkribus and open-source on Huggingface and GitHub. These models reach thousands of users every month. A remaining question is what significance this development has for the GLAM sector in general and the Swedish National Archives specifically.

To explore this issue, the paper presents findings from a survey conducted among professionals within Sweden’s GLAM sector. The results reveal a strong appreciation for the potential of citizen science and machine learning techniques, though many respondents lack practical experience with these methods. While some respondents view AI and citizen science as tools to improve efficiency, others emphasize co-creation and learning as key benefits. On one hand, these technologies are seen as offering significant advantages; on the other, respondents express concerns about the lack of resources, time, and knowledge required to integrate them into regular workflows. Additionally, we present findings from a second survey targeting researchers as archival users, which further investigates how archival institutions can evolve to meet changing user behaviors. The results from both surveys will be analyzed and compared.

Through these insights, the paper offers a comprehensive cultural and technological perspective on enhanced digital infrastructure, highlighting the role of collaborative AI models in shaping the future of archival access and use.

Asko Kauppinen, Pille Pruulman-Vengerfeldt, Åsa Harvard-Maare, Oskar Aspman
Malmö University, Sweden

Dreaming of perfect data work – three creative methods to understand the dreams, nightmares and practices of working with data in a museum

Thursday, 06/Mar/2025 3:20pm - 3:40pm
ID: 192 / Session SP 06: 5
Short paper (abstract) | 15-minute presentation with a 5-minute Q&A
Keywords: Creative methods, data, data work, emotional work, museums

This short paper outlines three creative methods experiments conducted in a medium-sized Nordic museum, which, in addition to a museum collection also has archival, educational, and library roles, as well as the role of caring for the built historical environment. The work-in-progress short paper focuses on the questions of data as an object of study, but instead of looking at what data itself does, we focus on looking how people working with data think, feel and act around the data. Our preliminary findings indicate that data emerges, precisely, as an issue dreams, nightmares and practices of 1) dream of perfect data work, 2) nightmare of fragmentation, perceived lack of routines, rituals and leadership; 3) but above all, an issue of interpersonal unresolved emotional labour.

We are “coming out of the ivory tower” to work together with the museum to understand the challenges they face when working with the data. Through creative experiments, we elicit individual reflections and group reflections about the data at work, data for the institution, and data for the sector. The presentation outlines three creative methods experiments conducted in a medium-sized Nordic museum, which has, in addition to a museum collection also archival, building-preservation and library roles, which, according to our participants tend to form silos of data work, rather than a cross-disciplinary knowledge institution. The three creative exercises invited a group of selected museum professionals (n=17) to understand the role of data in their work.

The three methods included: first, a sink-or-float ship that would carry data and data work to a direction, second, a play on classic Nordic mythology with the world-tree Yggdrasil, and, third, a play on monster-creation/collage work where the participants were invited to build a collage to represent a helpful/working monster Kratt, and address a letter of request to the mythical creature to help with data work. All of these three methods had a strong visual component, and they worked with the cultural and creative imaginaries of the participants. The research team includes two visual artists who created the visual prompts for the discussions. Through the creative visual expressions, we were able to tap into people’s understanding (or lack thereof) of shared cultural imaginary for expressing otherwise invisible, and hard to capture data work. When working with the collage monster Kratt, the participants were given art magazines, children’s comics, technology, culinary and home decor magazines which provided plenty of materials for cutting and creating visual for representing the monsters who were supposed to assist in the complicated data work.

From the initial discussions, it was clear that the words of data, information and knowledge were often used interchangeably and while in the analytical sense, the differences are very relevant, distinguishing between different elements in the practice was not considered important by our participants.

This is a work in progress, where our interpretation of digital humanities is not looking at what digital tools and technologies can do to enhance our research, but rather, we look at how different ideas of the digital intervene with the work practices of a traditional humanities-oriented organization.

Ernesta Kazakėnaitė[1], **Justina Mandravickaitė**[2]
[1]Vilnius University, Lithuania; [2]Vytautas Magnus University, Lithuania

DH in practise: how do philology students analyse texts with DH tools

ID: 188
Poster and demo (abstract) with accompanying a 1-minute lightning talk
Keywords: DH teaching; Text analysis: Philology students

Introductory course to Digital Humanities in Vilnius University Faculty of Philology started to be taught five years ago as an elective course for bachelor students. It was a long-gestating idea that came from the students. The course focuses on hands-on activities and we promote student projects instead of an exam. No project topics are given, as to encourage students to try mixed methods, approaches and tools. Although we pay attention to the sound (speech analysis) and image (image/visual analysis), text analysis takes the major out of the course. With this in mind, we analysed how students of all five years chose text analysis related topics, approaches and tools.

Philology students are used to working with texts, know traditional approaches therefore our interest lied in exploring how or if they combine in their works traditional and computational methods, what textual genres have been popular as a research material, what types of tools students tended to use (e.g., web-based tools, visual programming-based tools, etc.). All this knowledge is relevant for adapting and personalising our DH course to make it even more accessible for students without a technological background.

Eugenia Kelbert Rudan[1,4], **Lucija Mandić**[1,2], **Sasha Rudan**[3]
[1]Instite of World Literature, Slovak Republic; [2]Institute of Slovenian Literature and Literary Studies ZRC SAZU; [3]Oslo University; [4]University of East Anglia

Nabokov in Colour: A Digital Analysis

Wednesday, 05/Mar/2025 3:00pm - 3:30pm
ID: 251 / Session LP 03: 4
Long paper (abstract) | 20-minute presentation with a 10-minute Q&A
Keywords: Nabokov, colour, interdisciplinary, synaesthesia, bilingualism

This paper analyses the function of colours in Vladimir Nabokov's novels with reference to his self-described colour alphabet, and uses it to identify a shift in his perception of colours in English and Russian, respectively. In this way, we observe the corpus of Nabokov's novels through the prism of his private perception of how colours and written language relate to each other on both the statistical and semantic levels.

It is well known that Nabokov had synaesthesia, a genetic condition, inherited from his mother, that compels the person to experience more than one sense simultaneously. We hypothesise that Nabokov's synaesthesia influenced his writing. In his autobiographical work *Conclusive Evidence* and the Russian and English editions of its revised version, *Speak, Memory*, he describes his synaesthesia in great detail, including the colour each letter of the alphabet evokes in his mind. Nabokov details his respective 'colour' alphabets in both English and Russian. The letters in different languages, he observes, produce somewhat different shades; the differences between the two alphabets and these elements of interference present a starting point for an analysis of Nabokov's oeuvre with a focus on his use of colours. We use his descriptions of his synaesthesia in both manuscript drafts and in three published versions of this passage to analyse and visualise the letter and consequently colour distributions across his bilingual oeuvre. A close analysis of the evolution of Nabokov's epithets that describe colour in both languages and of the relationship between colour epithets and other words in his novels complement this comprehensive analysis of the way Nabokov's synaesthesia informs his work and his bilingualism.

First, we calculate the relative frequencies of letters and the distribution of colours they represent, and compare the results with a reference corpus of English and Russian literature. Given the often vague and idiosyncratic nature of Nabokov's colour descriptions—where he frequently avoids conventional terms—we experiment with generative AI to visualise the colours he describes in Russian and English, offering a novel way to interpret his synaesthetic experiences. This allows us to effectively colour-map his works in ways that reveal how his perception of language is intertwined with visual imagery. Statistical analyses of the letter frequencies in Nabokov's novels show that their colour topographies differ measurably from chapter to chapter. This adds a new dimension to reading his work through uncovering the layers of meaning tied to his synaesthetic interpretations.

By using a range of computational methods, we then analyse a broader segment of Nabokov's oeuvre, allowing us to identify patterns and trends in his use of colour across different works and languages. This analysis not only broadens our perspective but also subsequently highlights specific passages that warrant closer examination.

In addition, we offer a broader analysis of the importance of colour in Nabokov's oeuvre. First, we build on the differences in Nabokov's synaesthetic perception in his respective languages to examine the relationship between his use of colours and his bilingualism. Psycholinguistic research suggests that colour perception can vary significantly among speakers of different languages. Considering existing research on colour perception between speakers of the languages Nabokov was fluent in (Russian, English, French and German) and bilingual speakers, we trace how the distribution of colour changes not only between his English and Russian works but also across time.

In the last step of our analysis, we employ an original digital tool, LitTerra, to graphically visualise each novel's vocabulary. This method of textual analysis is based on an original method of visualisation of individual words in a literary text as a network (we call this component Society of Words). Just like in actual reading, Society of Words gives us the ability to focus on thematic or grammatical aspects of the novel that interest us the most, such as literary characters, thematic clusters, parts of speech etc. and on the relationships between individual nodes. In each case, the network allows us to see which words are associated – forming a series throughout the work – with the given node representing, say, Humbert Humbert in "Lolita" and a given semantic group of words (in our case, colours) or a part of speech (in our case, adjectives). The basic connection represents the pattern whereby two words occur in close proximity of each other repeatedly throughout the text. In this way, we can model the ways in which words and concepts can be associated one with another in tangible ways that affect interpretation. This added perspective allows us to analyse, building on the first two stages of this research, how the use of different colours by Nabokov relates to characters and themes in his novels.

Finally, we use LitTerra to complement this multifaceted computational analysis with close readings of relevant passages. LitTerra integrates, among its other functions, stylometric analysis and its visualisation with close reading and machine alignment of multiple texts in different languages, enabling parallel comparative analysis. Integrating the different stages of our analysis within an overarching framework, we can seamlessly access the passages where colour epithets are identified by computational methods as particularly significant, either in semantic terms or in relation to Nabokov's synaesthesia or bilingualism. This allows us to augment our findings with a semantic analysis of the use of colour in Nabokov's bilingual fiction. In this way, distant reading and close reading approaches are used in a complementary fashion to create a comprehensive overview of how a semantic category that is especially important to the writer unfolds in his work in relation to different characters and topics. This research adds to the study of Nabokov, of bilingual literature and of stylometry, presents a synthetic approach to computational literary analysis and, finally, opens up a window onto ways in which literature can be affected and informed by cognitive and psycholinguistics factors.

Ambika Dawn Kirkland, Jens Edlund
KTH Royal Institute of Technology, Sweden

Tabletop roleplaying games as a tool for interaction research

ID: 271 / Poster Session 2: 26
Poster and demo (full-text) with accompanying a 1-minute lightning talk
Keywords: Spontaneous conversation, interaction, roleplaying games, experimental design

Once a relatively obscure phenomenon, tabletop roleplaying games (TTRPGs) have exploded into mainstream popularity in the past decades. To date, much of the research involving TTRPGs has focused on roleplaying games as a creative medium, their role in constructing narratives and identities, and the motivations and characteristics of people who play them. However, TTRPGs also hold potential as a tool for investigating research questions across a range of disciplines, including those which do not directly consider games and gamers as objects of study. The medium of roleplaying provides a backdrop for many different types of multiparty interactions with a diversity of speech acts, all contained in the same context: e.g., narratives, collaborative problem-solving, explanations, suggestions, reasoning, instructions, discussion of moral dilemmas, argumentation, self-directed utterances, negotiations, questions, etc. The semi-structured nature of roleplaying games also makes it possible to organically incorporate specific tasks or other content into an interaction, or to present multiple parties to the conversation in the form of non-player characters (NPCs) animated by the game leader. TTRPGs can be run in person, via video chat platforms, or even in text form, making them adaptable to a variety of experimental paradigms. Using a number of interactions we have collected from roleplaying games run via Zoom, we present examples of the types of data which can be gleaned from TTRPG sessions and some potential methods for designing TTRPG-based studies in digital humanities topics.

Lars Kjær, Anders Klindt Myrvoll
The Royal Danish Library, Denmark

Workshop on cultural heritage data mining

Tuesday, 04/Mar/2025 2:00pm - 6:00pm
ID: 130 / WS14
Half-day conference-themed workshop

Keywords: AI, text and data mining, cultural heritage data, no-code, visual programming, workhsop, introduction

Are you fascinated by AI and text and data mining? Want to get insight into fundamental concepts in the field of machine learning? And are you interested in finding ways to use cultural heritage data as material in your Digital Humanities courses?

Then take part in this workshop and learn about this using a user-friendly software called Orange. Orange is a comprehensive, component-based software suite for machine learning and data mining, developed at Bioinformatics Laboratory, Faculty of Computer and Information Science, University of Ljubljana, Slovenia, together with open source community (Demsar J, Curk T, Erjavec A, Gorup C, Hocevar T, Milutinovic M, Mozina M, Polajnar M, Toplak M, Staric A, Stajdohar M, Umek L, Zagar L, Zbontar J, Zitnik M, Zupan B , 2013).

The advantage of Orange is that it is a no-code software that handles many of the same tasks and processes that are integrated parts of the data science that typically is applied to different fields of studies within both natural, social and humanistic science.

Orange can be used by researchers and employees in the GLAM sector. If you need to understand and evaluate results made by digital methods, then Orange is great for learning about AI and text and data mining. If you wish to introduce Digital Humanities in a classroom, then Orange can be used instead of Python or R for multiple data science tasks. If you need to build datasets and/or experiment with new approaches to a digital collection, then Orange can be used to sort, classify, analyse, categories, and retrieve data.

Target audience: University researchers and employees in the GLAM sector.

Expected learning outcome: On this workshop you can learn about:

- Orange's documentation
- Orange's interface
- Importing data into Orange
- Preprocessing text; preparing data for natural language processing tasks
- Embedding; transformation of images and text into vectors
- Classifying in order to analyse and interpret data
- Visualisation of data

Format: The half-day workshop is based on active participation and switches between short talks and practical solving of small tasks.

Technical requirements and data: To participate you got to bring your own computer, and I would also ask you to have downloaded and installed Orange on your computer before the workshop begin. Click here to go to the download page. If you wish to work on your own data in form of image files, text files, csv files you are very welcome to do so, if you do not have your own data, then I will have data available for you to use.

Bibliography

Demsar J, Curk T, Erjavec A, Gorup C, Hocevar T, Milutinovic M, Mozina M, Polajnar M, Toplak M, Staric A, Stajdohar M, Umek L, Zagar L, Zbontar J, Zitnik M, Zupan B (2013) Orange: Data Mining Toolbox in Python, Journal of Machine Learning Research 14(Aug): 2349−2353.

Andres Kõnno[1], **Kais Allkivi**[2], **Jaagup Kippar**[2], **Taavi Kamarik**[2], **Herman Petrov**[2]
[1]University of Tallinn, Baltic Film, Media and Arts School, Estonia; [2]University of Tallinn, Institute of Digital Technologies, Estonia

The role of metadata in artistic research: Reflections from the BFM student film archive

Thursday, 06/Mar/2025 3:20pm - 3:40pm
ID: 209 / Session SP 05: 5
Short paper (abstract) | 15-minute presentation with a 5-minute Q&A
Keywords: linked data, semantic search, keyword extraction, artistic research, film archives

This presentation reflects over the preliminary outcomes of the FilmEU+ subproject WIRE, which entails 8 film schools across Europe. The aim of this subproject is to create a cross-referenced student film archive that a) employs linked data and b) is built on the re-use of existing metadata and the generation of new metadata.

In this presentation, we intend to showcase the creation of metadata from the film archive of Baltic Film, Media and Arts school (BFM). BFM student film database consists of approximately 2,000 student films from the 1990's up to present day, out of which around 600 are made visible in the Estonian Film Database (EFIS). The quality and amount of metadata attributed to those films vary greatly, which means that there is a need for additional layers of metadata.

As we look at the artistic research that can be done via studying film metadata, there are three main categories: 1) descriptive metadata (director, actors, genre, year of production, film festivals, etc); 2) technical metadata (video resolution, sound design, visual effects, etc) and 3) content-based metadata (topics, tropes, visual motifs, narrative structures, etc). Our focus is a) on the study of content and b) on combining the descriptive and technical aspects of metadata.

We propose a contemporary organisation of semantic research that is based on the study of film texts and audiovisual data with the means of linguistic algorithms, in the wide array of possible outcomes of these methods: named entity recognition (NER), syntactic analysis, keyword extraction, and topic modelling. We suggest that the existing amount of metadata can be enriched with a new layer describing potential semantic connections which could facilitate the work of artistic researchers.

As we look at the potential outcomes of metadata-based artistic research of film archives, we discuss five following aspects:

1) The analysis of film data with the means of linked data and semantic search encourages researchers to explore new relationships and interpretations that contribute to a richer understanding of cinema as an art form.

Employing linked data in artistic research allows film metadata to be connected with other cultural or historical databases, creating a network of interconnected information. For example, linking film data with databases of historical events or literary works can provide richer context and deeper insights into the cultural and artistic significance of a film. It also means that, using linked data, researchers have to build ontologies that define relationships between different entities in film, such as characters, themes, settings, and historical references. This allows for more complex queries and deeper exploration of how themes or motifs are used across different films. Most importantly – this is where artistic research meets the work of historians, sociologists, semiotics, and cultural science.

2) Unlike keyword-based search, semantic search understands the context and relationships between terms, making it easier to find relevant information. For instance, if a researcher is exploring the topic of "gender alienation" in cinema, semantic search can identify films that deal with related concepts, even if those films do not explicitly use the word "gender" or "alienation". By using semantic search, researchers can discover less obvious connections between films or topics, finding hidden links between directors, genres, or narrative elements.

Also, the retrieval of metadata and content-based data becomes more efficient, enabling researchers to focus on the analysis and interpretation rather than spending time searching for relevant data manually.

3) Creating knowledge graphs and/or maps of connections between films based on shared topics, directors, or historical contexts can reveal new insights. This approach could offer a more intuitive understanding of data, helping to identify underlying patterns or relationships that may not be immediately evident through numerical or textual analysis.

4) The principal focus on historical characteristics of film data would allow us to interpret their changes over time (the past 30 years in case of the BFM database) – especially how films represent societal norms, cultural values, and historical events. Researchers can also explore how specific genres or directors contribute to cultural discourse and artistic movements via examining topics like identity, memory, and power.

5) The interoperability of different datasets might also include the data of film audiences. Via studying connections, this type of artistic research could embrace the subjectivity of interpretation, focusing not only on quantitative data but also on how data evokes meaning. For example, researchers can explore how audiences interpret and respond to films, or how cultural topics are perceived and transformed in various contexts.

Krister Kruusmaa[1,2], **Peeter Tinits**[1,3], **Laura Nemvalts**[1]
[1]National Library of Estonia; [2]Tallinn University, Estonia; [3]University of Tartu, Estonia

Unlocking Cultural Insights: Curating the Estonian National Bibliography for Research

Thursday, 06/Mar/2025 9:30am - 10:00am
ID: 237 / Session LP 05: 2
Long paper (abstract) | 20-minute presentation with a 10-minute Q&A
Keywords: data curation, library data, national bibliographies

National bibliographies are collections of data that compile information on printed publications associated with a particular country or ethnic group. While traditionally used for cataloging, these bibliographies are increasingly being repurposed as datasets for cultural historical studies, contributing to the field of bibliographic data science (e.g., Lahti et al. 2019a, Umerle et al. 2023). Such datasets can be leveraged to explore topics like historical intellectual networks (Hill et al. 2019) or the canonicity and popularity of written works (Tolonen et al. 2021). However, a significant challenge lies in transforming catalog data into standardized, scalable, and research-ready datasets (Lahti et al. 2019b, Mäkelä et al. 2020). Cataloging practices, designed for precision rather than analytical workflows, often vary over time and rely on outdated formats like MARC, making data refinement essential.

In response to this challenge, we developed a curated dataset from the Estonian National Bibliography (ENB), comprising over 300,000 book records. Our approach builds upon the broader movement toward open science, FAIR data, and reproducible research, aligning with ongoing efforts in the GLAM (galleries, libraries, archives, and museums) sector to enhance the accessibility of digital collections. Traditionally, bibliographic data has been stored in formats optimized for internal cataloging, with limited computational usability. By applying an open-source, modular, and scalable workflow, we transform this dataset into a structured, research-ready resource that can be explored using both basic spreadsheet software and advanced data analysis tools.

The curation process involved selecting ~50 fields of high cultural and research relevance, including author information, publication locations, publishers, languages, keywords, physical details, etc.. We employed a range of data processing techniques, including extraction, cleaning, harmonization, and enrichment. For instance, historical place names were standardized using external geographical databases, publisher names were harmonized through rule-based and vector similarity methods, and author data was linked to external authority files like VIAF and Wikidata. These improvements significantly enhance the dataset's suitability for computational analysis.

Beyond dataset creation, our work highlights the potential of national bibliographies as valuable resources for cultural history research. We illustrate this through case studies demonstrating the dataset’s utility in analyzing publication trends, mapping intellectual networks, and examining shifts in translation practices. The dataset reveals long-term trends in book publishing, illustrating how political changes influence literary production. Due to the extensive detail and broad coverage of the ENB, the curated dataset is highly adaptable for diverse case studies across various research interests.

By advocating for curated datasets within GLAM institutions, we underscore the importance of collaborative efforts between researchers and memory institutions. The expertise of catalogers, combined with computational methods, offers new opportunities for bibliographic data science. Our work demonstrates that curated datasets not only enhance accessibility but also foster interdisciplinary research by making national bibliographies more adaptable to contemporary digital scholarship. We emphasize that our approach is fully open-source, modular, and scalable, making it potentially applicable to other national bibliographies and similar datasets. In this paper, we detail our workflow, present the curated dataset, and explore its implications for bibliographic research, emphasizing how structured data can drive new perspectives in cultural and literary studies.

Bibliography

DATASET: Kruusmaa, K., Tinits, P., & Nemvalts, L. (2025). Curated Estonian National Bibliography - books (Version 3) [Data set]. National Library of Estonia. https://doi.org/10.5281/zenodo.14708287

PIPELINE: https://github.com/RaRa-digiLab/enb-curator

Hill, M. J., Vaara, V., Säily, T., Lahti, L., & Tolonen, M. (2019). Reconstructing intellectual networks: From the ESTC’s bibliographic metadata to historical material. Proceedings of the Digital Humanities in the Nordic Countries.

Lahti, L., Marjanen, J., Roivainen, H., & Tolonen, M. (2019a). Bibliographic Data Science and the History of the Book (c. 1500–1800). Cataloging & Classification Quarterly, 57(1), 5-23.

Lahti, L., Vaara, V., Marjanen, J., & Tolonen, M. (2019b). Best practices in bibliographic data science. In J. H. Jantunen, S. Brunni, N. Kunnas, S. Palviainen, & K. Västi (Eds.), Proceedings of the Research Data And Humanities (RDHUM) 2019 Conference: Data, Methods And Tools (pp. 57-65). (Studia humaniora Ouluensia; Vol. 17). University of Oulu.

Mäkelä, E., Lagus, K., Lahti, L., Säily, T., Tolonen, M., Hämäläinen, M., Kaislaniemi, S. & Nevalainen, T. (2020). 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

Tolonen, M., Roivainen, H., Marjanen, J., & Lahti, L. (2019). Scaling up bibliographic data science. 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 (pp. 450-456). (CEUR Workshop Proceedings; No. 2364). CEUR-WS.org.

Tolonen, M., Hill, M. J., Ijaz, A., Vaara, V., & Lahti, L. (2021). Examining the Early Modern Canon: The English Short Title Catalogue and Large-Scale Patterns of Cultural Production. In I. Baird (Ed.), Data Visualization in Enlightenment Literature and Culture (pp. 63-119). Palgrave Macmillan. https://doi.org/10.1007/978-3-030-54913-8_3

Kara Kuebart
University Bonn, Germany

Statebuliding in Jülich-Kleve-Berg. An example for a developing tax state within the 16th century empire

ID: 265 / **S01: 7**
Doctoral Consortium
Keywords: state building, taxation, 16th century, HTR, GIS

While theories on early modern state building are plentiful, they are oftentimes lacking in their scientific foundation. One reason for this is that case studies on early modern state building, especially in the area of taxation and fiscal capacity, require extensive research through often unsorted archival material in various unfamiliar handwritings. The bad accessibility, only partial preservation and improvised methods of early-modern 'bookkeeping' make the creation of comprehensive statistics on early modern state formation a very costly and time-consuming research endeavour.

My doctoral project has the aim of conducting such a case study by utilizing modern computational and AI-driven methods to facilitate the otherwise very tedious research process.

The United Duchies of Jülich-Kleve-Berg are an excellent object for such a case study. Their early modern finances have hitherto not been researched (as is the case for most smaller states in Europe), and they can be described as a typical worldly composite state within the empire of the 16th century. In addition, they are composed of five separate territories or 'states', if this term can be applied in the 16th century: The duchies of Jülich, Kleve and Berg, the Counties of Mark and Ravensberg and the Lordship of Ravenstein. Each territory had its own estates where the local nobility and the cities were represented. Both the balance of power and the administrative structures differed between these 'states', making it possible to compare these structures and their effects on fiscal capacity.

They also offer the possibility to determine whether a decentralized structure is a disadvantage in state development, a theory that has been intensively argued over in theoretical literature. I wish to determine whether the composite state suffered from "free riding" effects - as is the case in the Habsburg territories of the era -, whether the differences between the territories hindered the administration, and whether administrative structures have become more similar over time. I also wish to provide statistics on tax income, tax types, the assertiveness of tax collection, ducal and state debts, debtors, interest rates and fiscal capacity per capita. The intention is to make research results comparable to other case studies in the future.

My sources are reports on the financial situation in the five states of the United Duchies as well as in smaller administrative units like the "Ämter", reports on debts and debitors, protocols from sessions of the estates, decrees of the estates, tax matricula and tax lists from some of the "Ämter". The latter are used to verify if taxes granted in the matricula match the ones actually collected.

I digitize paper sources and utilize HTR methods (Transkribus) for the digitization of the text itself. Extensive corrections of these transcriptions are necessary, since models perform poorly on 16th century everyday handwriting. The resulting Ground Truth is used for training a more specialized model.

Text mining methods are used to help in the classification and prioritization of sources in my research.

Along with more traditional methods of source criticism and analysis I use statistical methods to analyse and interpret information on tax income and debt, and also geocode this information in order to trace the origin of this income locally. GIS methods are used to visualize the spatial dimension of taxation and determine the influence of a ruler over his territories on maps.

Results I have achieved so far point towards the decentral structure of this composite state effecting its progress positively. While administrative structures converged in their development to become more similar, the estates were keen on involving themselves in the composite state's political development while at the same time preserving their independence from one another. This becomes particularly apparent during the reign of the last duke, who was considered mentally ill. During this period, members of the estates involved themselves actively in the government of the composite state.

Kara Kuebart[1], Christian Schultze[2], Felix Selgert[1]
[1]University Bonn, Germany (History Department); [2]University Bonn, Germany (Computer Sciences)

Article based topic modelling on re-processed historical newspapers

Thursday, 06/Mar/2025 3:20pm - 3:40pm
ID: 259 / Session SP 07: 5
Short paper (abstract) | 15-minute presentation with a 5-minute Q&A
Keywords: image processing, newspapers, article segmentation, OCR, text mining

Historical newspapers contain a wealth of information about past societies. They provide information about the spatial occurrence of events, about contemporary perceptions of social and economic change, and allow tracing of cultural change.

Even though newspaper portals are an essential source for historians and other disciplines interested in history, such as economics, their potential has not yet been fully realized, due to several limitations. One of these is that the mass utilization of newspaper data is often limited to a keyword search, which usually only covers the entire page and does not discriminate between articles. Therefore, the joint occurrence of two or more search terms is recorded for the page, not the article, and information retrieval is thus still very imprecise. Furthermore, OCR quality and differences in language usage can create considerable bias in search results. AI driven textual data quality improvements have been proposed to amend this, but in most cases render unsatisfactory results. This can be amended by processing the pages anew from the images, using Deep learning algorithms that recognize the layout of a newspaper page and capture the text at the article level.

The purpose of this paper is to present such an AI pipeline and to demonstrate its implementation on a use case.

Our pipeline allows scholars of the humanities to apply advanced NLP-methods, including document and text embedding techniques, fine-tuning large language models, but also smaller, more focused analyses using text mining methods like sentiment analysis or named entity recognition. Also, Topic Modeling can now be used to sort articles into categories, facilitating the creation of a topic-specific corpora and reducing (albeit not eliminating) bias created by OCR errors and linguistic differences.

It is based on our Chronicling Germany Dataset, which contains over 700 pages of layout, baseline and text annotation from more than 20 different German-language Fraktur newspapers from 1700 to 1924. This also includes more than 50 pages of ads, which are of particular interest due to their more complex layout, irregular fonts and illustrations. We aim to publish a lengthy discussion of the pipeline and dataset in a computer science journal later this year (working paper: https://arxiv.org/pdf/2401.16845v3).

In short, we use a Convolutional Neural Network to detect page layout and recognize different layout elements, namely background, paragraph, table, heading and separator. For each found box containing legible text (this excludes background, table, separator and image classes), we perform a baseline detection (CNN U-Net). Then, a modified Kraken OCR is used to recognize the text of each line. We reach an OCR accuracy of 98% and an overall pipeline accuracy of 97% (average Levenshtein distance per word of 0.03). Results are saved in .xml format and can be exported to .csv with each text block as a row with text and metadata. Connecting multiple text blocks belonging to one article is a still outstanding project.

As a case study to demonstrate the use as well as potential research applications for the pipeline, we use topic modelling techniques to track various topics of political and societal importance in the “Kölnische Zeitung”, one of Germany’s most important newspapers in the 19th and first half of the 20th century.

We use a dataset of 44 000 newspaper issues (around 280 000 pages) of the “Kölnische Zeitung” from the period of 1803-1930, acquired from the newspaper portal “zeitpunkt.nrw” and processed through our aforementioned pipeline.

We will analyze political sentiment towards other European countries using the “international affairs” section each issue contains, which we extract by topic model. In this section, each paragraph begins with the fat-printed name of the country concerned, followed by a (usually biased) report on recent events there. Using sentiment and co-occurrence analysis on this data, we can track the newspapers opinion on national awakenings in the Baltic and the later ensuing wars of independence, but also its changing sentiment towards England, which at the beginning of the 19th century was seen as an ally. We can use newspapers to track how this viewpoint shifted.

The analysis is performed using both “leet-topic”, an NLP-based Transformer model for topic-modeling, and LDA topic-modeling executed using “gensim”. The code for sentiment analysis is our own lexicon-based implementation.

We would also like to highlight that our pipeline generalizes well to other languages set in Fraktur font and can thus be used on a variety of European newspapers. Particularly in the Baltic area and in Scandinavia Fraktur has persisted well into the 19th, sometimes into the 20th century, where our pipeline can serve for article separation and OCR on German-language and non-German sources, although retraining the OCR section of the pipeline on corresponding language data is recommended. For layout recognition, our dataset can also generalize to Antiqua fonts. It can also be used to pre-train new models on different fonts.

Bibliography

Beach, Brian; Hanlon, Willam Walker: Historical newspaper data. A researcher’s guide, in: Explorations in Economic History 90 (2023), vol. 101541.

Blevins, Cameron: Space, Nation, and the Triumph of Region. A View of the World from Houston, in: Journal of American History 101, 1 (2014), pp. 122-147.

Burchardt, Jørgen: Source criticism, bias, and representativeness in the digital age. A case study of digitized newspaper archives, in: DHNB (2024).

Burchardt, Jørgen: Are Searches in OCR-generated Archives Trustworthy? An Analysis of Digital Newspaper Archives, in: Jahrbuch für Wirtschaftsgeschichte 61, 1 (2023), pp. 31-54.

Carlson, Jacob; Bryan, Tom; Dell, Melissa: Efficient ocr for building a diverse digital history. ArXiv preprint arXiv:2304.02737, 2023.

Clausner, Christian; Papadopoulis, Christos; Pletschacher, Stefan, Antonacopoulos, Apostolos: The ENP Image and Ground Truth Dataset of Historical Newspapers. In: International Conference on Document Analysis and Recognition 13 (2015), pp. 931-935.

Davis, Brian et al.: End-to-end document recognition and understanding with dessurt. In: European Conference on Computer Vision (2022), pp. 280–296.

Dell, Melissa et al.: American stories: A large-scale structured text dataset of historical us newspapers, in: Advances in Neural Information Processing Systems, 36, 2024.

Ferrara, Andreas; Ha, Joung Yeob; Walsh, Randall: Using digitized newspapers to address measurement error in historical data. In: Journal of Economic History 84, 1 (2024), pp. 271-306.

Kiessling, Benjamin: The Kraken OCR system. In: https://kraken.re, accessed on: 20.01.2025.

Kodym, Oldrich; Hradiš, Michal: Page Layout Analysis System forUnconstrained Historic Documents. In: International Conference on Document Analysis 2021, vol. 12822.

Liebl, Bernhard; Burghardt, Manuel: From historical newspapers to machine-readable data. The origami ocr pipeline, in: Proceedings of the Workshop on Computational Humanities Research 2020, pp. 351-373.

Nikolaidou, Konstantina; Seuret, Mathias et al.: A survey of historical document image datasets. In: nternational Journal on Document Analysis and Recognition 25 (2022), pp. 305–338.

Oberbichler, Sarah; Pfanzelter, Eva: Topic-specific corpus building. A step towards a representative newspaper corpus on the topic of return migration using text mining methods, in: Journal of Digital History 1 (2021).

Ares Oliveira, Sofia; Seguin, Benoit; Kaplan, Frederic: dhsegment. A generic deep-learning approach for document segmentation. In: International Conference on Frontiers in Handwriting Recognition 16 (2018), pp. 7-12.

Smith, Ray: An overview of the tesseract ocr engine. In: International Conference on Document Analysis and Recognition 9 (2007), pp. 629-633.

Ströbel, Philip; Clematide, Simon: Improving ocr of black letter in historical newspapers. The unreasonable effectiveness of htr models on low-resolution images, in: Digital Humanities 2019.

Hanna-Mari Kristiina Kupari[1], Timo Korkiakangas[2], Veronika Laippala[1]
[1]University of Turku, Finland; [2]University of Helsinki, Finland

Building the Penitentiary Document Corpus (PeDoCo) for NLP: Balancing Data Complexity and Uniform Data Structure

Thursday, 06/Mar/2025 2:20pm - 2:40pm
ID: 144 / Session SP 07: 2
Short paper (full-text) | 15-minute presentation with a 5-minute Q&A
Keywords: corpus creation, digital humanities, Latin, TEI-XML, historical linguistics

This paper describes the process of creating a TEI XML corpus of late medieval Latin documents for NLP tasks from books in PDF format. The documents of the Apostolic Penitentiary (a tribunal of the Catholic Church responsible for granting absolutions, dispensations, and indulgences) have been originally published as printed books. For the purposes of this corpus, they were derived from PDF files used for proofreading before printing. These editions, containing 1,511 documents and 211,398 words, are designed by and for human scholars engaged in close reading. As a result, they encode implicit semantic information through typographical features such as page layout and italics, which are sometimes inconsistent. Although human readers, equipped with holistic understanding, can interpret such variations, NLP tools require unambiguous, text-only input. We report in detail the process of transforming the PDF editions into a structured, machine-readable, and openly accessible corpus. Our approach combines a rule-based workflow using regular expressions with close reading and manual corrections. Such conversion procedures, which are highly timeconsuming and require in-depth knowledge of medieval Latin philology and manuscript studies, are regrettably seldom made explicit, despite their vital role in ensuring the reproducibility and scalability of research.

Kyrre Kverndokk
University of Bergen, Norway

SAMLA: Some reflections on the development of a trans-institutional online tradition archive

ID: 302 / WS10: 4
Tradition Archives Meet Digital Humanities II
Keywords: Digital archive, trans-institutional archive, intermedial translation, sustainability

In September 2024, the online tradition archive SAMLA was launched. The digital archive includes the main collections from the three largest tradition archives in Norway. These tradition archives contain a rich diversity of cultural expressions and practices of majority, minority and indigenous cultures and spans genres and fields such as folktales, ballads, legends, beliefs, magic, food and craft traditions, and more. The online archive is the outcome of the digitization project SAMLA (2020–2025), which aims to make the folkloric source material digitally accessible for research, education, and business purposes. This presentation will focus on the challenges of developing a sustainable online, trans-institutional archive.

Digitization of archives may be regarded as intermedial translation processes, that raise a series of questions. Based on the experience from developing SAMLA, some of these questions will be this presentation: How to digitally coordinate the archive structure of three different archives. How to make the archive searchable, accessible and meaningful without an actively guiding archivist. How to administrate archive materials that are owned by different partner institutions. How to update the platform and maintain the archive after the project has ended. How to engage the audience also when the archive is not promoted by media appearances and special events.

Milda Kvizikevičiūtė

Martynas Mažvydas National Library Of Lithuania, Lithuania

eCulture: Advancing Accessibility and Curated Digital Content through a Unified Cultural Platform

ID: 304 / **WS02: 1**

Innovations and New Interactions through Digital Cultural Heritage in GLAM Sector

Keywords: Digital cultural heritage, Curated content, Accessibility, Audience engagement

The *eCulture* initiative, funded by the European Union's "New Generation Lithuania" plan, addresses critical challenges in the digitization, accessibility, and dissemination of cultural heritage in Lithuania. The GLAM (Galleries, Libraries, Archives, and Museums) sector has identified significant issues, including low levels of digitization and the insufficient or unappealing accessibility of digital and digitized cultural content. This project aims to create a unified digital platform that centralizes cultural content while emphasizing the development of curated products to enhance reuse, accessibility, and engagement for diverse user groups.

At the core of the platform is the creation of curated content tailored to the needs of specific audiences, including educators, researchers, children, cultural tourists, and entrepreneurs. Using design thinking methodology, the content curation process will undergo iterative testing and adaptation, ensuring the development of intuitive products that provide seamless access to cultural narratives and foster engagement and educational exploration. By integrating advanced technologies such as artificial intelligence (AI), 3D digitization, and semantic data models, *eCulture* transcends basic digitization processes to deliver thematically organized, high-quality digital resources.

The project will result in the creation of 100 curated content products of varying scale and technological sophistication, including collections, stories, blogs, maps, and virtual tours. This presentation outlines the challenges and opportunities involved in developing these curated products within the *eCulture* framework, addressing issues such as interoperability, accessibility, and audience-specific customization.

The project's partners include diverse institutions spanning cultural, artistic, archival, and media domains. These range from national repositories like the Lithuanian Integral Museum Information System (LIMIS) and Lithuanian Central State Archives to performing arts institutions such as the Lithuanian National Opera and Ballet Theatre, as well as inclusive organizations like the Lithuanian Audiosensory Library. This interdisciplinary collaboration ensures the platform reflects the richness of Lithuania's cultural heritage while meeting diverse user needs and fostering inclusivity, accessibility, and innovation.

By emphasizing the role of curated digital products, *eCulture* demonstrates how digital cultural heritage can bridge the gap between preservation and user engagement. This case study provides insights into the transformative potential of curated content for societal renewal, inclusivity, and innovation in cultural heritage practices.

Sandis Laime[1], **Sanita Reinsone**[2]
[1]Institute of Literature, Folklore and Art, University of Latvia; [2]Faculty of Humanities, University of Latvia

Garamantas.lv – Overview and Critical Reflections on a Decade of Digital Folklore Archiving

ID: 306 / **WS10: 2**
Tradition Archives Meet Digital Humanities II
Keywords: Archives of Latvian Folklore, data standardization, digitization, participatory approach, sustainability

The launch of the digital archive garamantas.lv by the Archives of Latvian Folklore (ALF) a decade ago marked a transformative phase in the ALF's development. Initially conceived as a platform for public access to digitized ALF collections, garamantas.lv has since become a central pillar of Latvia's digital humanities infrastructure, supporting participatory and collaborative practices alongside multimodal data exploration across multiple disciplines. This decade-long evolution has redefined both the conceptual and operational dimensions of ALF's archival work. The full digitization of its historical manuscript collections has been complemented by the rise of a dedicated community of volunteer transcribers, reflecting a fundamental shift toward participatory archiving. However, this expansion has also introduced complex methodological challenges, including issues of data standardization, curation, and the sustainability of digital research infrastructure. In this presentation, we will examine how the development of garamantas.lv has influenced the methodologies and practices of digital folklore archiving. We will explore how participatory approaches, public engagement, and scholarly curation can complement each other in a dynamic digital archive, while also addressing challenges related to infrastructure sustainability, methodological adaptation, and the long-term preservation of folklore collections.

Veronika Laippala, Petri Paju, Valtteri Skantsi, Hannu Salmi
University of Turku, Finland

Imagined Homelands: Analyzing the Finnish-language Press in North America 1876–1923 with Artificial Intelligence

ID: 173 / Poster Session 2: 4
Poster and demo (abstract) with accompanying a 1-minute lightning talk
Keywords: transnational press, historical newspapers, Finnish-American periodicals, immigrant discourses, genre detection, text reuse detection

New technologies - especially artificial intelligence - and the availability of newspapers and magazines in digital format have brought dramatic changes in the ways in which we can study historical materials. In particular, we can now ask novel questions that go beyond what is possible to answer through close reading and qualitative analysis.

In this poster, we present the project The Imagined Homelands and its first findings on how to study transnational press, in this case Finnish-language newspapers and magazines published in North America in 1876–1923. Our basic idea is that the textual construction of immigrant newspapers provides clues to how they imagined both past and present homelands. The immigrant (or ethnic) press took shape between cultures, publishing news from both Finland and North America, sometimes directly copying, but also reminiscing, reflecting, and imagining the conditions of the former homeland. The study of national cultures and nationalism has emphasized how nations are imagined and how images, meanings and symbols construct a sense of community. In this project, we will examine the perceptions of this community as shaped by the immigrant press.

The project explores what the methods of the digital humanities can tell us about the construction of the Finnish press in North America and its relationship to the former and present homeland. Specifically, we will apply text-reuse detection, and textual genre detection, as well as named entity recognition, enabling both close and distant reading of the substantial data.

Our data derives from the Finnish National Library's collection of Finnish-American periodicals containing a total of 226 newspaper and magazine titles since 1876. Publishers have been various political and spiritual associations of Finnish immigrants, but also local communities. The material is openly accessible digitally up to 1923. In addition, later publications can be used digitally in legal deposit libraries and also locally from microfilm.

In the poster, we will focus on first findings of two experiments: text segmentation and genre detection. The first task aims at identifying texts from the digitised material, where each document corresponds to a page in the newspaper. Texts are embedded within these pages, and sometimes also span several pages. Identifying these builds the basis for all the further analyses on the data. Genre detection, then, allows distinguishing between different kinds of texts published in the periodicals, ranging from news articles to letters and poetry. To what extent can these classes be automatically identified, and what kinds of central themes do the text classes feature? Experiments on both tasks are based on various deep learning approaches and the genre (register) identification tools developed originally for web genres (Henriksson et al. 2024, Skantsi & Laippala 2023). In particular, we test different GPT models and scenarios to achieve the best results.

Bibliography

Henriksson, E., Myntti, A., Eskelinen, A., Erten-Johansson, S., Hellström, S., & Laippala, V. (2024). Untangling the unrestricted web: Automatic identification of multilingual registers. arXiv. https://arxiv.org/abs/2406.19892

Skantsi V, Laippala V. Analyzing the unrestricted web: The Finnish corpus of online registers. Nordic Journal of Linguistics. Published online 2023:1-31. doi:10.1017/S0332586523000021

Lundell, P. Salmi, H., Edoff, E., Marjanen, J., Paju, P. & Rantala, H. (eds.) Information Flows across the Baltic Sea: Towards a Computational Approach to Media History. Mediehistoriskt arkiv, Lund 2023

Ilze Ļaksa-Timinska, Haralds Matulis, Sanita Reinsone
Institute of Literature, Folklore and Art of the University of Latvia

Crowdsourcing in Archives: A Case Study of the Latvian Pandemic Diaries Project

Wednesday, 05/Mar/2025 5:20pm - 5:40pm
ID: 191 / Session SP 01: 5
Short paper (abstract) | 15-minute presentation with a 5-minute Q&A
Keywords: crowdsourcing, citizen science, digital archives, Covid pandemic, diaries

Introduction

This paper focuses on crowdsourcing practices employed by the Archives of Latvian Folklore during the COVID-19 pandemic to gather digital diaries documenting the momentary and evolving experience of the pandemic unfolding. The paper touches upon challenges encountered by cultural heritage institutions when carrying out a rapid-response collecting campaign of autobiographical materials and outlines key insights for effectively managing such digitally mediated collecting efforts.

Crowdsourcing the current moment: challenges of digital archives

As the first wave of the COVID-19 pandemic unfolded and lockdowns were imposed in most parts of the world in March 2020 many cultural heritage institutions felt the need and even pressure to collect pandemic testimonies.[1] Now, several years later, reflecting on the practices of documenting the COVID-19 pandemic experiences as part of intangible cultural heritage, it is clear that the pandemic has been recorded by many institutions and individuals in physical, printed, and digital formats. From an institutional perspective, including that of museums, libraries, and archives – the primary custodians of cultural memory – it is evident that, compared to previous historical health crises, this pandemic was marked by an overwhelming abundance of information.[2]

The initiatives that are most accessible to the public, however, are predominantly collaborative archival projects,[3] which aim to preserve a wide range of perspectives documenting these historic times.[4] The way archives function has shifted in recent decades – a new paradigm that reflects the evolving realities of the digital environment, where archivists transition from being exclusive experts behind institutional walls to becoming mentors, facilitators, and coaches who engage with communities to promote archiving as a participatory process shared widely across society.[5] The initiative of the Latvian Pandemic Diaries Collection, is discussed in the paper, can also be categorized as this latter type of project.

The Pandemic Diaries Collection is part of a continuing effort by the Institute of Literature, Folklore and Art of the University of Latvia (ILFA) to establish an effective methodological and technological framework for cultural heritage crowdsourcing, suitable for various contexts and the collecting and processing of diverse material types. It is anchored in the Digital Archive of Latvian Folklore, which has been developed since 2014 and is based on the core principles of openness, participation, and sharing. The digital archive primarily houses the Archives of Latvian Folklore (ALF) collections, including manuscripts, photographs, audio and video recordings, as well as born-digital collections; it is available open access via website: garamantas.lv.

In the case of the Pandemic Diaries Project, the availability of readily accessible digital infrastructure for crowdsourcing was a key factor enabling the smooth organization of a rapid-response collecting campaign. Rapid-response strategies, aimed at quickly mobilizing and gathering data during critical moments, have proven essential in times of crisis. The COVID-19 pandemic highlighted the acute need for digital tools and methodologies to support such data collection, as rapid-response initiatives often emerged as the most effective, and at times the only viable, ways of documenting the pandemic period.

In the creation of the ALF Pandemic Diaries Collection, the approach to documenting the pandemic combined a rapid-response collection strategy with a crowdsourcing methodology, building on the existing digital infrastructure. An important step was an open call, inviting as many individuals as possible to participate in the documentation effort to create the collection, which would preserve documented experiences of pandemic for future generations. An open call went beyond a simple invitation to participate – it became a way to share information and build strong relationships with media, who were eager to report on the Pandemic Diaries project and encourage others to get involved. The promise to document the current moment and the need to engage participants in a digital community of Pandemic diarists prompted project organizers to make records publicly available in the digital archive garamantas.lv as soon as possible (usually on the same day or the next day as the records are submitted). As the project's success and increased visibility attracted more participants, the growing influx of diaries placed significant pressure on the available human resources needed to process and publish them on a daily basis. This showed the limitations of methodologies chosen and led to reconsidering the resources available and how to reshape the ongoing project, balancing between managing the expectations of participants, keeping up the best practices of collecting and documenting, and the workload of project organizers.

Over the course of the first year of the pandemic (March 2020 to April 2021), 2334 diary entries were archived (1 day = 1 entry), written by 238 different authors. All records were dated, accompanied by known metadata (information about the author, place of writing), and locations mentioned in the texts were geo-tagged.

This paper, proposed for DHNB2025 conference, will present a comprehensive overview of the methodologies and technical frameworks used in the digital archiving and development of the pandemic diary collection. The second part of the presentation will focus on a quantitative analysis of the corpus of pandemic diary texts, providing statistical insights into writing patterns, submission frequency, metadata analysis, spatial mapping of geo-tagged entries, and thematic exploration through topic modeling.

Conclusion

The pandemic period was challenging for many, including heritage institutions. It became evident that new digital agendas were essential, and that documenting the pandemic under the constraints of social restrictions posed a significant challenge. Institutions aiming to document this historical moment had to adapt quickly and innovate to record the momentous changes impacting society. The Latvian Pandemic Diaries initiative provides an exemplary case study in this regard. An examination of the diaries collection created during the pandemic demonstrates the diverse possibilities for analysis afforded by such collections. They hold significant value, capturing not only personal experiences but also serving as a broader cultural archive that reflects the diverse social and emotional landscape of a turbulent period.

Bibliography

[1] Kosciejew, Mark, ‘Remembering COVID-19; or, a Duty to Document the Coronavirus Pandemic’, in: IFLA Journal 48:1 (2022), 20-32.

[2] Jones, Esyllt W. et al., 'Remembering Is a Form of Honouring: Preserving the COVID-19 Archival Record', in: FACETS 6 (2021), 545-568, 550.

[3] Noordegraaf, Julia et al., ‘Microscopic Views on a Global Pandemic: Social and Cultural Effects of the COVID-19 Pandemic as Documented in Two Dutch Community Archives’, in: Journal of Open Humanities Data 7 (2021), Article 16.

[4] Zumthurm, Tizian, 'Crowdsourced COVID-19 Collections: A Brief Overview', in: International Public History 4:1 (2021), 77-83.

[5] Cook, Terry, ‘Evidence, Memory, Identity, and Community: Four Shifting Archival Paradigms’, in: Arch. Sci. 13 (2013), 95-120, 114.

Rafael Leal[1], Annastiina Ahola[1], Eero Hyvönen[1,2]
[1]Aalto University, Finland; [2]University of Helsinki, Finland

Enriching Cultural Heritage Metadata with LLMs and Knowledge Graphs: Case Finnish Named Entity Linking

Thursday, 06/Mar/2025 10:00am - 10:30am
ID: 210 / Session LP 05: 3
Long paper (abstract) | 20-minute presentation with a 10-minute Q&A
Keywords: Large language models, Named Entity Recognition, Named Entity Linking, Linked data, Knowledge graph

This paper presents work on using Large Language Models (LLM) to disambiguate Named Entity Linking candidates, which is meant for enriching the metadata of textual documents by linking them to Knowledge Graphs. We propose a zero-shot classification method that has similarities with Retrieval-Augmented Generation (RAG), and discuss 1) a prototype web service and 2) a user interface on top of it that allows for human intervention when making final disambiguation decisions, since this cannot be reliably carried out in automatic fashion due to errors and hallucinations of LLM-based tools. The focus of this work is on Finnish texts, so our methods take into account the particularities of this highly inflectional language and the resources available for processing it. This paper presents promising preliminary evaluation results of the system, suggesting feasibility of the methods and tools presented: our named entity lemmatizer achieved an accuracy of 96.5% on our test dataset, and a local LLM of the Llama family was able to find the correct linking candidate in 16 out of 17 examples. GPT-4 achieved 100% accuracy in linking using both standard text and YAML format.

1. Introduction and Motivation
Much of the data that could be used in Digital Humanities (DH) research is available only in unstructured textual form. Information extraction is then needed for creating metadata based on Knowledge Organization Systems (KOS) and Knowledge Graphs (KG) (Martinez-Rodriguez, Hogan, and Lopez-Arevalo 2020), publishing Linked Data Services, and building applications on top of them, such as the Sampo systems (Hyvönen 2022). For example, in our work on publishing the plenary session speeches of the Parliament of Finland as Linked Open Data (LOD), the speeches had to be linked to various domain-specific ontologies based on named entities (people, places, organizations, etc.), keyword resources, and a library classification system (Tamper et al. 2022). A fundamental task here is Named Entity Recognition (NER) and Linking (NEL). This paper addresses the question of how Large Language Models (LLM) can be exploited for the task where semantic disambiguation is a key challenge. This work is focused on Finnish texts, and we discuss some of the pitfalls that incur when carrying out natural language processing in this language.
It is important to notice that one of our aims is to be aware of the computational power needed to run the system. In many cases, we chose smaller, specialized local models when a call to an external LLM could be carried out. Moreover, we also strive to use open tools wherever possible.

2. Related works
Traditionally, Named Entity Recognition and Named Entity Liking have been separated into different tasks, although newer works, capitalizing on the breakthroughs of deep neural networks, focus on end-to-end linking. For example, Logeswaran et al. (2019) created a zero-shot linking system focused on addressing domain-shifting, without the need for gazeteers, while Ayoola et al. (2022) performs entity detection and disambiguation in a single forward pass. With the advent of LLMs, the field of Retrieval-Augmented Generation (Lewis et al. 2020) expanded dramatically, as summarized in Fan et al. (2024). The interface between LLMs an knowledge graphs can be considered a subdomain of RAG, and there are many recent works in this field: Baek, Aji, and Saffari (2023) created a LLM-based framework to answer questions by verbalizing triples related to entities in knowledge graphs, while Edge et al. (2024) used abstractive summarization over an entire corpus to find answers to global questions.
Regarding works about the Finnish language, Mäkelä (2014) described a web-based tool for annotating texts based on linked data, which included named entities. Luoma et al. (2020) released a corpus and tool for Finnish Named Entity Recognition, expanding on Ruokolainen et al. (2020). This work has been used for example for pseudonymization of court documents (Oksanen et al. 2019) as well as studying Parliamentary data (Tamper et al. 2022). However, as far as we are aware, there are no previous works using neural network techniques for entity linking in Finnish.

3. Disambiguation and Linking
NER in Finnish as a practical task has been well-served since the publication of the command line tool by the TurkuNLP group in 2020 (Luoma et al. 2020). However, the additional, more difficult task of disambiguating and linking entities (NEL) to external knowledge graphs (KG) has not been as prominent, despite the advent of LLMs. In this paper, we utilize a classification method that bears similarities with RAG in order to disambiguate and link entities in Finnish texts to linked data resources.
Our approach follows the traditional entity linking procedure, divided into three steps: 1) entity recognition; 2) candidate generation; 3) disambiguation and linking. Our system recognizes entities via third-party tools, such as the aforementioned NER tool. Although LLMs could be used for this purpose, their ratio of computing power demands to accuracy can be significantly worse, and their idiosyncrasies can make it more difficult to extract the entities themselves. The recognized entities are then lemmatized (changed to their basic forms) and lexically matched to a knowledge graph in order to find suitable candidates. There are around 15 nominal cases in Finnish, which makes it a significantly harder language to lemmatize than more analytic ones, for example English, which tend to use separate words to indicate case. Since word-by-word lemmatization does not produce satisfactory results for Finnish named entities, we fine-tuned Finnish-NLP/t5-small-nl24-finnish [1], a T5 generative model containing around 260 million parameters, with a dataset of Wikipedia internal links of our creation. The training set contains around 1 million pairs of named entity surface forms (alongside their context) and their basic form (based on their page names), and it will be released as open data. The candidates are generated from a local database containing Wikipedia articles and related Wikidata, which have long been used as primary targets for named entity linking (Mihalcea and Csomai 2007). The generation is done via lexical matching, which means that word forms, rather than their meaning, are used to find similarities. This technique might be more fragile than alternatives based on vectorization, since it uses word forms instead of context for candidate generation. It also requires a wealth of alternative labels to be included in the

databases, so that entities can also be found via surrogate names. However, it allows for easy integration of any number of databases and knowledge graphs, since it does not require pre-processing or fine-tuning, and dispenses with tracking changes to them.

For many named entities, the number of plausible candidates can be heuristically shrunk to one, for example matching detected entity type to possible candidates, which bypasses the need for further processing. Otherwise, the candidates are presented to the LLM, which is tasked with deciding which one is the most suitable, based on the context in which the entity appears in the text. This step has some characteristics in common with RAG, since both use external retrieval to enhance prompting and capitalize on the emergent capabilities of LLMs to learn in-context (Chan et al. 2022). However, here the LLM works as a zero-shot classifier rather than a generative model: the information presented to the LLM represent a strict narrowing of generative output rather than contextual information to draw upon. This restriction lessens the tendency of LLMs to hallucinate. Furthermore, we ask the model to spell out the reasoning behind its choices, since this is known to improve generation (Kojima et al. 2022).

Our aim is to link to candidates in any number of databases. In order to achieve this KG-agnostic status, the information related to the candidates extracted from the knowledge graphs should be presented to the LLM in an appropriate format. Standard text descriptions are better suited than knowledge graphs for LLMs to reason upon, which is expected due to the nature of LLM pre-training. However, extended textual descriptions are not common in knowledge bases, so as an alternative test setup we retrieve Wikidata properties for the candidates and format them as YAML, with property labels in Finnish whenever possible, or in English as a backup. This format was chosen for the ease of conversion from RDF graphs and its minimal markup, which translates into fewer LLM tokens.

4. A Tool for Automatic Annotation

The purpose of our research is to create not only a strong baseline for automatic annotation of Finnish documents but also a user interface that supports its application. As such, the entity linking tool is being integrated into a front-end interface that could enable users to seamlessly revise and correct the named entity links found in a document and then save the metadata for further use in DH research and applications. Such a tool can be used for example in cases where the annotations are critically important, such as in dealing with legal documents or parliamentary data [2] (Hyvönen et al. 2024). The user would then be able to upload the text to the tool, review and edit entity links proposed by the tool, and the save the corrected metadata in a fashion similar to the pseudonymization tool ANOPPI (Oksanen et al. 2019), previously created by our research group.

5. Results

5.1. Named entity lemmatization

Our named entity lemmatizer obtained an accuracy of 96.5% on 10K test examples from the Wikipedia internal links dataset. Since many characters have to be added to the model's tokenizer in order to correctly process foreign-language entities, we also fine-tuned a multilingual version of T5 [3], which include such characters during training, but this model only achieved an accuracy of around 83%. It seems that the multilingual version does not incorporate Finnish grammatical rules well enough for it to succeed in this task.

5.2. Entity disambiguation

Two LLM models were tested for entity disambiguation: Llama-3-8B-Instruct (Touvron et al. 2023) and GPT-4 (OpenAI et al. 2024). A total of 17 disambiguation examples were used for this task. These examples were automatically extracted from a Finnish Wikinews dataset that we have labelled, and are all related to Wikipedia disambiguation pages. This test sample includes, among others, locations (Gothenburg, Odessa (Texas), Kaduna), organizations (Esso, Citroën), and persons (John Roberts). There is an average of 5.7 candidates for each entity in this test dataset, with a minimum of 3.

Two different kinds of candidate descriptions were tested: 1) Standard text, using the introductory part of the respective Wikipedia articles for each candidate, and 2) Wikidata contents formatted as YAML, as mentioned above, in which case only 5 examples were used. Llama-3-8B-Instruct correctly identified 16 out of 17 cases using standard text, but it failed in all 5 cases using YAML. The model understands the task and returns one of the candidates, but it cannot identify the correct one. GPT-4, on the other hand, obtained an accuracy of 100% in both tasks.

6. Discussion

Our main objective in working on this project is to increase the digital presence of Finnish, making the existing metadata-processing tools for this language more robust and full-featured. However, working with a language which is so peripheral even in an European context is challenging, which can be seen in practice in the absence of tools and resources such as open datasets.

This paper describes an entity linking system in Finnish, which capitalizes on existing tools and models in order to extract named entities, generate candidates for them and choose the best candidate for linking. It showed that existing LLMs are capable of dealing with Finnish text satisfactorily, and the most powerful ones are even able to parse predicates in YAML with full accuracy. These results suggest the feasibility of the methods and tools presented here. In the future, we will test larger local LLMs, such as Llama-3.1-70B-Instruct, to probe their capabilities of disambiguating candidates with properties formatted as YAML. This would allow us to utilize knowledge graphs that do not store standard text descriptions while avoiding committing to a single LLM. Furthermore, our web interface for editing links will be finalized and integrated with our system. The system will be open-sourced in its entirety, alongside our dataset of fully-linked Wikinews texts as well as the lemmatization dataset.

Endnotes

[1] https://huggingface.co/Finnish-NLP/t5-small-nl24-finnish

[2] Our group has released, among others, the LawSampo (https://lakisampo.fi/) and ParliamentSampo (https://parlamenttisampo.fi/) data services and semantic portals which use linked data.

[3] google/mt5-small

Ayoola, Tom, Shubhi Tyagi, Joseph Fisher, Christos Christodoulopoulos, and Andrea Pierleoni. 2022. “ReFinED: An Efficient Zero-shot-capable Approach to End-to-End Entity Linking.” In Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Industry Track, 209–220. NAACL-HLT 2022. Hybrid: Seattle, Washington + Online: Association for Computational Linguistics, July. https://doi.org/10.18653/v1/2022.naacl-industry.24.

Baek, Jinheon, Alham Aji, and Amir Saffari. 2023. “Knowledge-Augmented Language Model Prompting for Zero-Shot Knowledge Graph Question Answering.” In Proceedings of the First Workshop on Matching From Unstructured and Structured Data (MATCHING 2023), edited by Estevam Hruschka, Tom Mitchell, Sajjadur Rahman, Dunja Mladenić, and Marko Grobelnik, 70–98. MATCHING 2023. Toronto, ON, Canada: Association for Computational Linguistics, July. https://doi.org/10.18653/v1/2023.matching-1.7.

Chan, Stephanie, Adam Santoro, Andrew Lampinen, Jane Wang, Aaditya Singh, Pierre Richemond, James McClelland, and Felix Hill. 2022. “Data Distributional Properties Drive Emergent In-Context Learning in Transformers.” In Advances in Neural Information Processing Systems, edited by S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, 35:18878–18891. Curran Associates, Inc.

Edge, Darren, Ha Trinh, Newman Cheng, Joshua Bradley, Alex Chao, Apurva Mody, Steven Truitt, and Jonathan Larson. 2024. “From Local to Global: A Graph RAG Approach to Query-Focused Summarization.” Pre-published, April 24, 2024. https://doi.org/10.48550/arXiv.2404.16130. arXiv: 2404.16130 [cs].

Fan, Wenqi, Yujuan Ding, Liangbo Ning, Shijie Wang, Hengyun Li, Dawei Yin, Tat-Seng Chua, and Qing Li. 2024. “A Survey on RAG Meeting LLMs: Towards Retrieval-Augmented Large Language Models.” In Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, 6491–6501. KDD ’24. New York, NY, USA: Association for Computing Machinery, August 24, 2024. isbn: 9798400704901. https://doi.org/10.1145/3637528.3671470.

Hyvönen, Eero. 2022. “Digital Humanities on the Semantic Web: Sampo Model and Portal Series.” Semantic Web – Interoperability, Usability, Applicability 14 (4): 729–744. https://doi.org/10.3233/SW-190386.

Hyvönen, Eero, Laura Sinikallio, Petri Leskinen, Senka Drobac, Rafael Leal, Matti La Mela, Jouni Tuominen, Henna Poikkimäki, and Heikki Rantala. 2024. “Publishing and Using Parliamentary Linked Data on the Semantic Web: ParliamentSampo System for Parliament of Finland.” Accepted. https://www.semantic-web-journal.net/system/files/swj3605.pdf.

Kojima, Takeshi, Shixiang (Shane) Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. “Large Language Models Are Zero-Shot Reasoners.” In Advances in Neural Information Processing Systems, edited by S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, 35:22199–22213. Curran Associates, Inc.

Lewis, Patrick, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, et al. 2020. “Retrieval-augmented generation for knowledge-intensive NLP tasks.” In Proceedings of the 34th International Conference on Neural Information Processing Systems. NIPS ’20. , Vancouver, BC, Canada, Curran Associates Inc. isbn: 9781713829546.

Logeswaran, Lajanugen, Ming-Wei Chang, Kenton Lee, Kristina Toutanova, Jacob Devlin, and Honglak Lee. 2019. “Zero-Shot Entity Linking by Reading Entity Descriptions.” In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, edited by Anna Korhonen, David Traum, and Lluís Màrquez, 3449–3460. ACL 2019. Florence, Italy: Association for Computational Linguistics, July. https://doi.org/10.18653/v1/P19-1335.

Luoma, Jouni, Miika Oinonen, Maria Pyykönen, Veronika Laippala, and Sampo Pyysalo. 2020. “A Broad-coverage Corpus for Finnish Named Entity Recognition.” In Proceedings of the 12th Language Resources and Evaluation Conference, 4615–4624. LREC 2020. Marseille, France: European Language Resources Association, May. isbn: 979-10-95546-34-4, accessed November 30, 2021. https://aclanthology.org/2020.lrec-1.567.

Mäkelä, Eetu. 2014. “Combining a REST Lexical Analysis Web Service with SPARQL for Mashup Semantic Annotation from Text.” In The Semantic Web: ESWC 2014 Satellite Events, edited by Valentina Presutti, Eva Blomqvist, Raphael Troncy, Harald Sack, Ioannis Papadakis, and Anna Tordai, 424–428. Cham: Springer International Publishing. isbn: 978-3-319-11955-7. https://doi.org/10.1007/978-3-319-11955-7_60.

Martinez-Rodriguez, Jose L., Aidan Hogan, and Ivan Lopez-Arevalo. 2020. “Information Extraction Meets the Semantic Web: A Survey.” Semantic Web – Interoperability, Usability, Applicability 11 (2): 255–335.

Mihalcea, Rada, and Andras Csomai. 2007. “Wikify!: Linking Documents to Encyclopedic Knowledge.” In Proceedings of the Sixteenth ACM Conference on Conference on Information and Knowledge Management, 233–242. CIKM07: Conference on Information and Knowledge Management. Lisbon Portugal: ACM, November 6, 2007. isbn: 978-1-59593-803-9. https://doi.org/10.1145/1321440.1321475.

Oksanen, Arttu, Minna Tamper, Jouni Tuominen, Aki Hietanen, and Eero Hyvönen. 2019. “Anoppi: A Pseudonymization Service for Finnish Court Documents.” In Legal Knowledge and Information Systems. JURIX 2019: The Thirty-Second Annual Conference, edited by M. Araszkiewicz and V. Rodríguez-Doncel, 251–254. IOS Press, December. isbn: 978-1-64368048-4. https://doi.org/10.3233/FAIA190335.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, et al. 2024. GPT-4 Technical Report. arXiv: 2303.08774 [cs.CL].

Ruokolainen, Teemu, Pekka Kauppinen, Miikka Silfverberg, and Krister Lindén. 2020. “A Finnish News Corpus for Named

Entity Recognition." Language Resources and Evaluation 54, no. 1 (March 1, 2020): 247–272. issn: 1574-0218. https://doi.org/10.1007/s10579-019-09471-7.

Tamper, Minna, Rafael Leal, Laura Sinikallio, Petri Leskinen, Jouni Tuominen, and Eero Hyvönen. 2022. "Extracting Knowledge from Parliamentary Debates for Studying Political Culture and Language." In Proceedings of the 1st International Workshop on Knowledge Graph Generation From Text and the 1st International Workshop on Modular Knowledge co-located with 19th Extended Semantic Conference (ESWC 2022), edited by Sanju Tiwari, Nandana Mihindukulasooriya,

Francesco Osborne, Dimitris Kontokostas, Jennifer D'Souza, and Mayank Kejriwal, 3184:70–79. International Workshop on Knowledge Graph Generation from Text (TEXT2KG 2022). CEUR WS, May. http://ceur-ws.org/Vol-3184/TEXT2KG_Paper_5.pdf.

Touvron, Hugo, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, et al. 2023. LLaMA: Open and Efficient Foundation Language Models. arXiv: 2302.13971 [cs.CL].

Andreas Lenander Ægidius
Royal Danish Library, Denmark

Web Archives For Music Research

Thursday, 06/Mar/2025 2:40pm - 3:00pm
ID: 116 / Session SP 05: 3
Short paper (abstract) | 15-minute presentation with a 5-minute Q&A
Keywords: advocacy, digital music, musicology, user engagement, interviews

A European national library has set a strategic goal to make more of its cultural heritage materials accessible and engaging for researchers by 2024. In this paper, we present findings from an advocacy initiative targeted at researchers at national universities in music-related fields.

Danish universities engage in a wide range of music-related research, spanning from musicology and composition to music psychology and music technology. Researchers explore everything from musical perception and the cultural significance of music to the development of new digital tools for music production. Additionally, there is a focus on the role of music in society, including its influence on identity and social interaction. Collaborations between disciplines such as psychology, sociology, and arts contribute to a deeper understanding of music's complex nature and its significance in people's lives. Music collections and the Web archive at the national library are documenting the roles music plays in historic and contemporary societies (Royal Danish Library, n.d.-a; Royal Danish Library, n.d.-b).

The national web archive provides primary sources and contextual information relevant to music researchers as they engage with our music collections. However, there is room for improvement in the connection between these collections and our understanding of user needs. Ethnomusicology researchers argue that archivists and archives should play a more proactive role: not only can existing collections be better utilised, and usefully and meaningfully disseminated, but the way that archiving is done can also be transformed better to meet the needs of the cultural lives of those communities who find existing collections of particular value (Landau & Topp Fargion, 2012). Digital humanities supported by larger amounts of available data challenges the music related fields to reflect on their data analysis and use of computational methods (Egan, 2021).

Reports by Healy et al. (2022) and Healy & Byrne (2023) explore the challenges researchers face when using web archives, highlighting the ongoing need to examine the skills, tools, and methods associated with web archiving. Collaborations between those who create the data and those who want to use the data is proving to be a proactive solution for increasing scholarly engagement with web archives (Healy & Byrne, 2023 p. vii). Additionally, the sounds of the web—from MIDI to streaming—are an integral part of its history, yet this aspect is often overlooked by tools like the Internet Archive's Wayback Machine (Morris, 2019).

Through semi-structured interviews with curators, librarians, and music researchers at universities, we identify key barriers to access and user requirements for improved utilization of web archival resources. Web archiving is often suffused in technical details of maintaining the archival process of running and planning web crawls that produce the Web archive. Recent political economical prioritizations have minimized the structural framework that support advisory meetings between curators, researchers, and the commercial sectors. As a result, curators, who are preoccupied with technical tasks, receive less input from researchers. This lack of communication leaves little time for curators to stay updated on the evolving needs and desires of various research fields that could benefit from library collections. Our advocacy initiative also allows us to summarize current research trends as feedback for the web curators. In conclusion, we describe how the web curators processed our findings into suggestions for updates and refinements to web crawling strategies and the built-in tools in the SolrWayBack installation.

Bibliography

Egan (Pádraig Mac Aodhgáin), P. (2021). Insider or outsider? Exploring some digital challenges in ethnomusicology. Interdisciplinary Science Reviews, 46(4), 477–500. https://doi.org/10.1080/03080188.2021.1872144

Healy, S., & Byrne, H. (2023). Scholarly Use of Web Archives Across Ireland: The Past, Present & Future(s) (WARCnet Special Reports). Aarhus University. https://cc.au.dk/fileadmin/dac/Projekter/WARCnet/Healy_Byrne_Scholarly_Use_01.pdf

Healy, S., Byrne, H., Schmid, K., Bingham, N., Holownia, O., Kurzmeier, M., & Jansma, R. (2022). Skills, Tools, and Knowledge Ecologies in Web Archive Research (WARCnet Special Reports). Aarhus University. https://cc.au.dk/fileadmin/dac/Projekter/WARCnet/Healy_et_al_Skills_Tools_and_Knowledge_Ecologies.pdf

Landau, C., & Topp Fargion, J. (2012). We're all Archivists Now: Towards a more Equitable Ethnomusicology. Ethnomusicology Forum, 21(2), 125–140. https://doi.org/10.1080/17411912.2012.690188

Morris, J. W. (2019). Hearing the Past: The Sonic Web from MIDI to Music Streaming. In N. Brügger & I. Milligan (Eds.), The SAGE Handbook of Web History (pp. 491–510). Sage.

Royal Danish Library (n.d.-a) How to find music. Web Resource. Retrieved from https://www.kb.dk/en/find-materials/guides/music. (Last visited October 7th 2024)

Royal Danish Library (n.d.-b) Netarkivet. Web Resource. Retrieved from https://www.kb.dk/en/find-materials/collections/netarkivet. (Last visited October 7th 2024)

Andreas Lenander Ægidius[1], **Masao OI**[2]
[1]Royal Danish Library, Denmark; [2]National Institutes for the Humanities, Japan

Collaborating to improve cultural heritage collections as data: analyzing Europeana and Japan Search

Thursday, 06/Mar/2025 3:40pm - 4:00pm
ID: 127 / Session SP 05: 6
Short paper (abstract) | 15-minute presentation with a 5-minute Q&A
Keywords: digital cultural heritage, Europeana, Japan Search, metadata, linked data

This paper reports on a Japanese and Danish collaborative project on digital cultural heritage infrastructures and data models (2024-2026). The aim of the projects is to compare how different approaches to providing access to digital cultural heritage (DCH) can learn from each other and improve their respective solutions. The project involves the following libraries and supportive institutions:

- *National Diet Library* in Japan wants to introduce links to DCH collections outside Japan into its Japan Search platform.
- Cabinet Office Government of Japan want to utilize international DCH collections in the future on Japan Search for educational and other purposes.
- *National Institutes for the Humanities* in Japan supports the development of inter-university research institutes that promote research in the humanities.
- *Royal Danish Library*, which serves as the national library, wants to improve access to its collections and provide better metadata and linked data. A significant goal is to publish links to its collections on the collaborative European initiative, the *Europeana* platform.
- *Europeana* empowers the cultural heritage sector in its digital transformation and advocates for better digital practices that support openness, transparency and reuse of digital cultural heritage.

The institutions support digital humanities education and to do so they have to work with their DCH collections as data (Candela & Gabriëls et al, 2023). Additionally, the institutions share a common goal to improve discoverability and connections with other related online collections across borders and alphabets. Thereby, the project contributes to advance a development of linked data in DCH collections that has been under way in the last ten years (Lambert & Southwick, 2013).

RQ1 general: what activities should the libraries, government institutions, and research institutions prioritize, preferably with low effort and high impact, which will improve access to their collections as data in Japan Search and Europeana?

RQ2 specific to data models: What are the most important fields in the data models and what requirements do they set for the improvement of metadata and the use of linked open data in an institutional setting?

Method

The paper reports on a series of workshops that bring together data scientists, librarians and cultural heritage curators, and researchers from national libraries and universities. In addition, the project participants manage and develop the workshops using collaborative literature reviews and desk research.

The workshops and the desk research employ a two-fold approach in a collaborative investigation of the Europeana and Japan Search portals:

1. from the outside-in we compare the collection metadata and linked data in the platforms user interfaces and how it is displayed (knowledge graphs).
2. from the inside-out we compare data models and publishing guides from Europeana (n.d.) and Japan Search (n.d.)

Alongside its mapping of the portals and their data models, the project draws on didactic considerations from existing research (e.g. Oi et al, 2022). Included in its focus on access, the project promotes the use of DCH in learning scenarios as alternatives to Google-based searches. Finding and using simple digital objects is fine, e.g. illustrations for a pupil's specific class presentation. However, in-depth exploration of historical source materials should not begin with Google Search and thereby risk enforcing the misconception that ‘Google is the internet’. Japan Search has the advantage of being available in the pupils’ native tongue and English. However, Oi et al. (2022) have shown that it is still labor-intensive to search for learning materials from a vast amount of information and consider how to utilize them in the classroom. In patrticular, when questions are developed from an international perspective based on domestic information, it is difficult to find and use resources from other countries. Answering RQ1 and RQ2 provides a basis for recommended ways towards increased use of Europeana and Japan Search in learning scenarios.

Preliminary findings

Our initial assumption was that the Royal Danish Library and Europeana have a lot to learn from Japan Search. One example of this is evident from the fact that only 4,6 percent of the Europeana database have the highest metadata tier while more than 50 percent have the lowest metadata tier (Metis, n.d.). The data models of Europeana and Japan Search support the Resource Description Format (RDF, 2014). However, it seems there are also comparable challenges shared by the parties involved, e.g. high barriers to entry when adopting the principles of Linked Open Data at GLAMs and international links between collection metadata. It seems the data models and the data mapping guidelines are still a well-meaning, yet tangled, network of documents that define, prime, guide, and provide case studies to a degree that will exhaust even the most hardened curators and developers. The institutions might succeed in publishing (links and descriptive metadata about) its collections. Still, we have found case where the follow up data management is lacking and users will find confusing metadata that suggest haphazard publishing activities.

Based on the comparative analysis and workshop outcomes, the research project will propose a set of best practices for cultural institutions seeking to implement or enhance their digital collections using the data models of Europeana and/or Japan Search. Our findings so far suggest the proposed best practices will emphasize the importance of user-centered design, metadata standardization, and the establishment of clear guidelines for linked data implementation. In conclusion, this paper will outline these best practices to be proposed and hope to discuss these with the digital humanities community.

Bibliography

Candela, G. Gabriëls, N. et al. (2023) A Checklist to Publish Collections as Data in GLAM Institutions. Retrieved from: https://arxiv.org/abs/2304.02603 (Last visited Oct. 11th 2024)

Europeana (n.d.) Web page. Retrieved from https://www.europeana.eu/en (last visited Oct. 9th 2024)

Europeana data model (n.d.) Web Page. Retrieved from: https://pro.europeana.eu/page/edm-documentation (last visited Oct. 10th 2024)

Japan Search data model (n.d.) Web page. Retrieved from: https://www.ndl.go.jp/jp/dlib/standards/jpsformat.html (last visited Oct. 10th 2024)

Lampert, Cory K., Southwick, Silvia B. (2013). Leading to Linking: Introducing Linked Data to Academic Library digital Collections, Journal of Library Metadata, 13:2-3,230-253.

Metis statistics dashboard for Europeana complete dataset (n.d.). Web page. Retrieved from: https://metis-statistics.europeana.eu/en/ (Last visited Oct. 9th 2024).

National Diet Library, Japan (n.d.) Web page. Retrieved from: https://www.ndl.go.jp/en/index.html (last visited Oct. 9th 2024)
National Institutes for the Humanities (n.d.) Web page. Retrieved from: https://www.nihu.jp/en (last visited Oct. 9th 2024)

Oi, M., Kim, B., Watanave, H. (2022). “S × UKILAM” Collaboration to Connect Local Digital Resources and School Education: Workshop and Archiving to Construct Network of “People” and “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. https://doi.org/10.1007/978-3-031-21756-2_10

RDF - Resource description format (2014) Web page. Retrieved from https://www.w3.org/RDF/ (last visited Oct. 11th 2024)

Erik Lenas, Viktoria Löfgren, Gabriel Borg, Morgan Svensson, Olof Karsvall
Swedish National Archives, Sweden

Evaluating Scalable Pipelines for Handwritten Text Recognition in Historical Swedish Archives

Thursday, 06/Mar/2025 2:40pm - 3:00pm
ID: 270 / Session SP 08: 3
Short paper (abstract) | 15-minute presentation with a 5-minute Q&A
Keywords: Handwritten Text Recognition, Open Source, Swedish National Archives

This paper investigates pipelines for large-scale Handwritten Text Recognition (HTR) in the context of the Swedish National Archives' ongoing, massive HTR project, where millions of scanned handwritten documents – spanning a wide variety of Swedish archives from 1600 to 1900 – are transformed into digital text. Our work assesses the accuracy and generalizability of different models and pipelines, establishing a strong baseline for future research and improvements.

Historical texts vary widely in handwriting styles, spelling conventions, and paper quality, posing significant challenges to model generalization across centuries and document domains. A core focus of our work is to evaluate the temporal and categorical adaptability of an HTR model trained on a large, diverse dataset. Fine-tuning experiments explore the effects of training with different amounts of period-specific data, from smaller subsets to larger collections. This approach allows us to assess the extent of performance improvement gained through fine-tuning and specialization of a base model, and how the size of the fine-tuning dataset impacts the character error rate (CER) and word error rate (WER) in text recognition.

We also present an extensive test set, consisting of gold-standard annotated samples from archives spanning the entire period for which the model was trained, but drawn from archives not included in the training set. This provides an indication of how well the model generalizes to new, unseen domains.

Evaluating an HTR pipeline requires more than assessing CER and WER on a text-line basis. A key focus of modern HTR pipelines is the segmentation of full pages into text regions and lines. Our assumption is that HTR pipeline models should not be evaluated in isolation using task-specific metrics alone; they should also be evaluated in an end-to-end manner, using page-level metrics that account for different reading orders. We examine variations in CER and WER across different domains, time periods, and document types, and assess these metrics in relation to factors such as the number of characters per line and lines per page. This analysis helps identify model outliers and suggests further adaptations and improvements to the pipeline.

Two main segmentation strategies are investigated and evaluated using the page-level metrics proposed. One strategy first segments pages into text regions and then further segments these regions into text lines. This nested approach is compared to a flat strategy, where region and line segmentation are performed simultaneously by the same model. We will later compare segmentation-based pipelines with emerging full-page HTR models, which aim to bypass the segmentation step entirely.

The results presented in this paper aim to contribute to the development of scalable HTR solutions, particularly in the context of large-scale digitization projects at cultural heritage institutions worldwide. Ultimately, the insights gained from our research will inform the advancement of Handwritten Text Recognition on a massive scale.

Evelina Liliequist, **Karin Danielsson**, **Jim Robertsson**
Umeå University, Sweden

Choreographing queerness: capturing motion of fluid identities

ID: 117 / Poster Session 2: 17
Poster and demo (abstract) with accompanying a 1-minute lightning talk
Keywords: Queer, AI, Motion capture

Previous studies about AI technology from a queer theoretical gaze have identified a number of problems and risks with AI systems, such as categorization and labelling, often in stereotypical ways, cementing normative performances of gender. Drawing on such insights, we have formed the interdisciplinary research group *Queer AI* at Humlab, Umeå University where we seek to combine our expertise in queer studies, informatics, digital humanities and computer science, with the aim to find new ways of building and implementing AI technology and systems in our society for an inclusive future.

In [author 1] et al. (2023) and [author 2] et al. (2023) we argue for the continued need and benefits of queer theory to make visible problems, risks and challenges with AI tech, and the increasing implementation of such tech in society, and the possibilities for adopting participatory approaches to increase users' involvement and strengthen their role during design of digital technologies and systems. By combining queer theoretical perspectives and participatory design approaches, we explore how fluid identities and performances can be expressed and represented in data, empowering contributors and transforming people from data subjects into being data creators on their own terms. In practice, we have conducted a pilot study where we explored how data retrieved from bodily movements can be both created, approached and understood by the data creators themselves, on their own terms. In the pilot study we used motion capture equipment to explore how data for AI-systems, for example in health care systems, chatbots or other interactive systems citizens may interact with on a daily basis, can be created in nuanced and diverse ways through various performances. By doing so our hope is to find new ways to create a more inclusive technology that is better adapted to be applied in a diversified reality.

Our work is part of a larger networking and community building, including seminars, conferences (research group: Queer AI), as well as a community reference meeting on Queering AI (event: WASP-HS).

Bibliography

Danielsson, K., Aler Tubella, 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). https://doi.org/10.4337/9781803928562.00061

Liliequist, E., Aler Tubella, A., Danielsson, K., & Cocq, C. (2023). Beyond the binary : queering AI for an inclusive future. Interactions, 30(3), 31–33. https://doi.org/10.1145/3590141

Queer AI: research group (2024) https://www.umu.se/en/research/groups/queer-perspectives-on-automated-facial-analysis-technology-queer-ai/

Queering AI – Community Reference Meeting WASP-HS (2024-10-02) https://wasp-hs.org/event/queering-ai/

Inna Lisniak[1], **Anna Aljanaki**[2], **Liudmyla Efremova**[3]
[1]Estonian Literary Museum, Estonia; [2]University of Tartu, Estonia; [3]Institute of Art Studies, Folkloristics and Ethnology of the National Academy of Sciences, Ukraine

A corpus of Ukrainian folk songs from Podillia

Thursday, 06/Mar/2025 2:40pm - 3:00pm
ID: 226 / Session SP 06: 3
Short paper (full-text) | 15-minute presentation with a 5-minute Q&A
Keywords: Computational music analysis, computational ethnomusicology, Ukrainian folk songs, a corpus of symbolic music

We present a corpus of 1310 symbolically encoded Ukrainian folk songs in musicXML format with lyrics and metadata, mainly from the West-Central and South-Western parts of Ukraine: Podillia. The corpus consists of three sections: ritual songs (calendar-ritual and family-ritual), non-ritual songs, and children's songs, which are three main types of traditional Ukrainian songs. These songs were collected by Ukrainian ethnomusicologists between 1914 and 2018. Some of the songs were recorded in audio format and transcribed to symbolic format by ethnomusicologists, but most songs (93%) do not have an audio recording in archive. We have substantially improved the quality of metadata and improved the searchability of the corpus by extracting details concerning song collection (year, location, ethnomusicologists names) from natural text description. We also extracted and present as a separate part of the corpus the lyrics of each song. We describe the properties of the corpus, comparing different song styles, and discuss new possibilities for using the Ukrainian song corpus in ethnomusicological research using computational methods.

Katarina Lučić
University of Bologna, Italy

From Art to WenDAng: Towards a Digital Archive for Chinese Contemporary Calligraphy

ID: 137 / S01: 4
Doctoral Consortium
Keywords: digital archive, semantic web, Linked Open Data, calligraphy, contemporary art

This paper presents the ongoing development of the WenDAng digital archive, designed to gather, preserve, analyse, and disseminate data collected by the WRITE project[1]. WRITE focuses on the emergence of new forms of calligraphy within contemporary Chinese art. Given the absence of specific ontologies for describing calligraphy, a key objective of the archive is to explore novel ways of representing contemporary Chinese calligraphy and to provide statistical insights addressing one of the project's central research questions: Can contemporary calligraphy still be considered calligraphy in traditional terms, and what are the similarities and differences between the two?

At the heart of the archive are four collections, each represented as a class corresponding to a distinct type of art: 1) visual arts, 2) decorative and applied arts, 3) performing arts, and 4) graffiti art. Another important feature of the archive is the "Calli-Writing" unit — a section of the artwork that contains some kind of calligraphy or any other written element. This unit, modeled as a distinct class, plays a pivotal role in comparing traditional and contemporary Chinese calligraphy. All these classes serve as the foundation for supporting domain experts' investigations, offering insights derived from three types of analysis: 1) artistic, 2) linguistic, and 3) socio-political-economic. To achieve this, the WRITE ontology is being modeled to represent the heterogeneous dataset, which includes not only artworks and "Calli-Writing" units, but also their contextual elements, such as artists, calligraphers, organisations, artistic collectives, places, exhibitions, literary works and others. The ontology builds on classes and properties from Wikidata to describe technical aspects of the artworks and their broader contexts. However, new ad-hoc classes and properties are being developed to represent calligraphy both as an artistic expression and as a set of signs.

Therefore, in the WRITE ontology, each class and its associated properties contribute to one or more types of analysis. Artistic analysis is supported by data on artworks and "Calli-Writing" units, covering aspects such as types, genres, materials, tools, colors, dimensions, subjects, and inspirations. Linguistic analysis focuses mainly on "Calli-Writing" units, literary works, critical texts, and exhibition catalogues, while socio-political-economic analysis deals primarily with graffiti art, decorative and applied arts, people, organisations, and exhibitions. For instance, by analysing the "Calli-Writing" unit from both artistic and linguistic perspectives, the archive can generate insights into how contemporary calligraphy relates to tradition. This is achieved by capturing data on materials (wdt:P186), tools (wdt:P2079), colours (of the objects wdt:P462 and background write:background-color), calligraphic lines (write:calligraphic-line), character forms (write:character-form), writing systems (write:writing-system), script styles (wdt:P9302), and, where applicable, the meaning of the written content (write:significance). Traditional calligraphy involves using a brush and black ink on white paper, with Chinese characters written in traditional script styles. By using the WRITE ontology, we can document and observe the contrasts in contemporary calligraphy, which may utilise diverse materials (not only ink on paper), tools (including digital ones), colours beyond black and white, new script styles, and other innovative features.

To enhance the representation of Chinese contemporary calligraphy, the archive is developing and storing controlled vocabularies[2] that include new taxonomies emerged through the domain experts' research. For now, these vocabularies include terms for script styles (22 terms), Chinese visual elements (34 terms), Chinese concepts (9 terms), and others. These vocabularies are another innovative aspect offered by the archive, especially if we compare them with other established vocabularies, such as Getty's Art & Architecture Thesaurus, that do not include contemporary script styles, Chinese visual elements present in the artworks nor other calligraphy-specific terms.

Data entry process is being facilitated through Omeka S[3], chosen for its user-friendly interface suited to domain experts. The dataset can be fully queried using Blazegraph[4], and preliminary queries conducted on over 200 artworks and 190 "Calli-Writing" units have already revealed notable distinctions between traditional and contemporary calligraphy. For instance, only half of the analysed artworks employed a brush, and merely one-third used a brush with ink and paper.

As the WRITE team's research continues, both the dataset and the ontology will evolve, incorporating new classes and properties. The WenDAng digital archive is expected to be completed by the end of 2026.

Footnotes

[1] The project "WRITE - New Forms of Calligraphy in China: A Contemporary Culture Mirror" has received funding from European Research Council (ERC) under the European Union's Horizon 2020 Research and Innovation Programme (GA n. 949645). For more information about the *WRITE* project, visit https://writecalligraphyproject.eu/.

[2] A specific module of Omeka S, Custom Vocab, is being used for creating controlled vocabularies.

[3] https://omeka.org/s/

[4] https://blazegraph.com/

Elisabeth Maria Magin
Kulturhistorisk museum, UiO, Norway

LaTeX Light: Introduction to LyX

ID: 230
Half-day conference-themed workshop
Keywords: LaTeX, writing and publishing academic works, layout, structured text

Despite the presence of computers in our daily lives, choosing the right tool for the right task can be tricky. This especially applies when we're already familiar with a specific tool for a specific task – like writing and publishing papers. Students in particular, but also many older colleagues, tend to use Microsoft Word or a similar word processor for the purpose, leading to a range of issues, not the least of which is compatibility. There are better tools available, like LaTeX, which offers a wide range of benefits over traditional Word documents, particularly for those submitting contributions to a variety of academic publications with different styles and requirements. LaTeX excels at compiling bibliographies and outputting them according to different stylesheets, meaning that resubmitting an article to a different journal only requires one change in the settings and a rerun for all of the references to be formatted according to the guidelines of the new journal. It further offers automated numbering of cross-references and handling of references, ensuring that table of contents, page numbers, bibliography and cross-references are always up-to-date without any manual work required. This is especially useful for those working with longer texts, where chapters and sections may have to be moved around during the writing process, and spares one the headache of having to manually update cross-references in the text.

Beyond these, LaTeX also offers other benefits. Using plain text and code to in, files can be edited on every computer regardless of access to proprietary software, rendering .tex-files completely portable and platform-indepedent; and since the visual output is decoupled from the content, it is also much easier to output the same document according to different stylesheets, for example using different fonts or page margins, than it is with Word.

But LaTeX can not only render the same document in different ways; it is also possible to create several different files out of the same basic document. This feature is especially useful if one happens to be writing worksheets for students or exam papers, or is a PhD student required to hand in a draft to one's supervisors, where some parts aren't quite finished yet. Keeping the student exam paper and the solution sheet, or the draft meant for one's supervisors versus the drafts one is still working on in different documents can easily lead to much confusion and wrong the versions of documents being sent to the wrong person.

But despite the many benefits of LaTeX, the switch from a word processor to a plain text editor with markup language can be intimidating and difficult to tackle by oneself. This is where LyX comes in. LyX is a graphical user interface that aims to make the transition from word processor to LaTeX easier by using many familiar buttons from word processors to enter the LaTeX code ”behind the scenes”, allowing newcomers to LaTeX to focus on getting the text written instead of trying to remember the right commands. However, the use of LaTeX in the background enables LyX users to make use of many of the features LaTeX offers – such as consistent formatting, easy cross-referencing, consistent bibliographies and creating different output documents from the same source document, thus keeping everything in the same space.

LaTeX Light: Introduction to LyX is meant for everyone looking for alternatives to Microsoft Word, whether student or established academic; everyone who has ever spent hours adapting the bibliography to a different journal's style or tried to stop the illustrations from completely messing up the formatting of the text. Over the course of four hours, this introductory course teaches up to 15 attendants how to write a structured text of any length – whether 5-page journal article or 15-chapter book – using LyX and compile it with LaTeX. In units of approximately 30 minutes, the course covers:
* The basic principles of “what you get is what you see” versus “what you get is what you want”
* The LyX interface
* Basic structuring in a LyX document (chapters, sections and subsections)
* Including images and tables in your document
* Bibliographies and cross-referencing
* Outputting different versions of the same document

To ensure that participants can do the practical exercises for each of the units, they should bring an unformatted paper draft (or some other text) with at least 3 sections, 3-5 subsections, 1 image and 1 table, which will be used as a training document. Participants are also required to install the latest version of either MikTeX or TeXLive (only ONE of them) and LyX before the workshop. All of these are freeware and available for Windows, Linux and Mac.

The course will be lead by Elisabeth Magin, who has given several introductions into LyX to a number of colleagues and students at the University of Oslo and conferences.

Bibliography

https://miktex.org/
https://www.tug.org/texlive/
https://www.lyx.org/

Elisabeth Maria Magin
Kulturhistorisk museum, UiO, Norway

Theoretically, It Could Be – Representing Runic Variation in Relational Databases

Wednesday, 05/Mar/2025 2:30pm - 3:00pm
ID: 247 / Session LP 02: 3
Long paper (abstract) | 20-minute presentation with a 10-minute Q&A
Keywords: runes, runology, ancient scripts, data structuring, entity modelling

Runes – a writing system most strongly connected to the Vikings in most people's perception and of particular interest to many, but in fact in use for more than 1000 years in some areas. Despite the public interest in the script, much of the information on the internet is outdated or comes from dubious parties; academic research into runes tends to remain within the scholarly sphere and unnoticed by the general public, not least because the field is very small and chronically underfunded. Using digital tools such as digital editions to publish reliable and up-to-date information on runic inscriptions seems like an obvious and comparatively cheap solution for runologists to convey their research to a bigger audience; yet, it is rarely an approach taken, and for good reasons. Despite the existence of the Unicode block Runic, runes are notoriously difficult to wrangle into digital form, and current approaches to using digital tools to both support research and aid dissemination of research leave much to be desired.

This paper presents the entity model developed as a result of the MSCA-funded project "From Stick To Screen – Digital Editions of Runic Inscriptions as Research Tools" and describes how it addresses a variety of challenges with encoding runes, using the process of deciphering ancient scripts as its basic framework. To show the applicability of the entity model and the approach presented, the paper further provides examples of how the data can be exported to be used in n-gram analyses of runes and their variations as they occur in the medieval inscriptions from Bergen, Oslo, Trondheim and Tønsberg. In doing so, the paper hopes to give other scholars working with different ancient scripts an example of how the methodological approaches can be successfully incorporated into an entity model, thus opening up new possibilities and angles for encoding and analysing corpora of ancient inscriptions.

One such challenge many ancient scripts (such as cuneiform or Khitan) share is the lack of standardisation across time and space, with many variations on the same basic shape being in use alongside each other. This also applies to runes; they are by no means "a writing system" – the first runic writing system, called the Older Futhark, has been remodelled at least two times. Even discounting post-Reformation use of runes as late as the 17th century AD, the Elder Futhark has spawned three closely related, but distinct systems: the Anglo-Saxon Futhorc, the Viking Age Younger Futhark and the Medieval Futhork. Consisting of one or more vertical staves and horizontal branches, runic writing displays a great deal of variability in shape. At times, more than 10 different variations of the same basic shape are in use, spread out over different geographic areas from Scandinavia to Britain to Greenland. This, as scholars of other ancient scripts know, is very common for all scripts that never underwent standardisation and were allowed to develop into regional and time-specific variants. Being able to track these variations is of great importance where the development of the script and the spread of (runic) literacy is concerned. Yet tracking them across time and space requires mechanical support.

Based on prior research into using relational databases as research tools in runology (Magin 2023, Data-Based Runes) and critical reflections on the Unicode code block Runic (Magin & Smith 2023, (R)Unicode), this paper shows how solutions for storing runes as characters based on a minimal character set using the Unicode block Runic can be implemented in a relational database. The entity model presented was developed alongside the process of deciphering such inscriptions. It is built specifically to allow ancient script scholars to edit their inscriptions character by character, adding comments to each as they go along, to render the process transparent. The paper will show how, taking into account some runologists are only interested in the runes as the characters used to create a written word, while others are interested in the specific visual traits, a complex entity model able to accommodate both levels of enquiry was developed (Magin in print, Runeninschriften aus Bergen). Further phenomena appearing in runic writing, such as different directions of writing (left-to-right, right-to-left), flipped runes or bindrunes (runic ligatures), are also stored in a standardised, and thus searchable, way.

Another challenge common for ancient scripts is the sheer breadth of readings presented by different scholars, especially for texts of particular interest or those only partially preserved. While databases such as Samnordisk runtextdatabas settle for presenting the most accepted and up-to-date interpretation, this paper demonstrates that it is possible to store conflicting readings and interpretations down to the level of descriptions of the visual traits of a single sign. This allows for direct comparison between the readings of different scholars and, in turn, enables others working with the dataset to judge the reliability of the single interpretation, or determine whether an inscription is potentially relevant for one's own research.

The ability to determine the reliability of a single interpretation is a vital component in conducting further studies on the material; in the case of this project, the aim was to create the basis to conduct diachronic and spatial case studies of the distribution of particular runic phenomena and personal names appearing in the medieval runic inscriptions from Bergen, Tønsberg, Oslo and Trondheim. By way of the special queries presented in the paper, the single runes in an inscription can be recombined into character strings. These can in turn be analysed with the help of software such as Excel, Python, R or SPSS, allowing the creation of n-grams and distribution maps for names. From these analyses, one can attempt to track the movement of people and/or goods between towns trading with each other, as Bergen, Oslo, Trondheim and Tønsberg were doing during the Middle Ages, and how this might have contributed to the spread of knowledge about different graphic variations of a specific rune or character. The paper presents examples of how this was done for specific cases, exemplifying what such analyses could contribute to our knowledge of the development of not just runes, but also other ancient scripts in the future.

Bibliography

Magin, Elisabeth Maria (2023). Data-Based Runes. Macrostudies on the Bryggen Runic Inscriptions. The Bryggen Papers Series, Bergen. DOI: https://doi.org/10.15845/bryggen.v100

Magin, Elisabeth Maria (in print). "Die Runeninschriften aus Bergen, Fonts und Unicode: Möglichkeiten der Digitalen Runologie". In: Gedenkschrift Klaus Düwel.

Magin, Elisabeth Maria & Smith, Marcus (2023). "(R)Unicode: Encoding and Sustainability Issues in Runology". In: Digital Humanities in the Nordic and Baltic Countries Publications 5, No. 1, p. 121-136. DOI: https://doi.org/10.5617/dhnbpub.10657

Mäkikalli Maija, Jokipii Ilkka, Joska Sanna
National Archives of Finland, Finland

Digitising cultural heritage of minority communities for equity and renewed engagement

Thursday, 06/Mar/2025 3:00pm - 3:20pm
ID: 169 / Session SP 08: 4
Short paper (abstract) | 15-minute presentation with a 5-minute Q&A

Keywords: minorities, digital cultural heritage, equity, community engagement, participatory approaches

The Digitisation of cultural heritage of minority communities for equity and renewed engagement (DIGICHer) project explores the legal, political, socio-economic and technological factors driving the digitisation of cultural heritage of minority groups. It is run by a multi-disciplinary consortium representing both academic and non-academic organisations working in the field of minorities' digital cultural heritage, digital humanities and law in five European countries. These partners, including Vilnius Gediminas Technical University (leader), Friedrich Schiller Universitat Jena, University of Lapland, Istituto Italiano di Studi Germanici, Europeana, National Archives of Finland, Jewish Heritage Network, Istituto Culturale Ladino, Lietuvos Inovaciju Centras, Network to Promote Linguistic Diversity and Time Machine, play a crucial role in the success of the project. Their broad expertise, diverse fields, and geographical coverage play an essential role in the project's activities and outcomes. This project has received funding from the European Union’s research and innovation programme Horizon Europe under Grant Agreement No. 101132481.

Based on their multi-disciplinary research, the project partners use co-creation approach to design a new and practical framework that promotes equity, inclusion and diversity in the representation of digital cultural content related to minoritised communities. The framework will be developed collaboratively with representatives from three minority groups in Europe: the Sámi, the Jewish and the Ladin people. The framework will be scalable and adaptable to other minorities as well. It will include recommendations for best practices while working with the digitised cultural heritage of minority communities. The framework will be helpful to a number of stakeholders involved in working with minority and their heritage, including policy and decision-makers and cultural institutions.

The project's collaboration with the Sámi communities is especially relevant in the Nordic context. As a project partner, the National Archives of Finland will participate in the development and validating the framework by conducting engagement and co-creation activities with members of the Sámi communities. We will organise a project for digitising archives in the holdings of the National Archives of Finland, including events and workshops discussing different aspects and best practices relevant to minorities' perspectives. We will also design evaluation of the Nuohtti search portal (www.nuohtti.com) which was launched in 2023. Nuohtti searches for digital Sámi archives in the collections of memory organisations in Europe. Nuohtti portal can be used in Northern Sámi, and it includes ethical guidelines for the respectful use of this often culturally sensitive digital material. (Häkkilä et. al. 2022; Moradi et. al. 2020; Tuominen et. al. 2023). Nuohtti is maintained in collaboration between the national archives of Norway, Sweden and Finland. The evaluation is done with all the actors involved in the service.

Our participatory approach process is designed to increase minorities' involvement in the digitisation and usage of their cultural heritage. It incorporates user-focused techniques, including design thinking and legal design, to ensure the communities validate the framework. As a result, the project creates knowledge-based recommendations for policy and decision-makers, ensuring that the needs and perspectives of the communities are at the forefront. DIGICHer collaborates between academic researchers and non-academic organisations and develops the societal relevance of digital humanities, making the audience feel that their input is valued and integral to the project's success.

This intervention will introduce and share learnings from the work of the project as detailed above.

For more information on the DIGICHer, see https://www.digicher-project.eu/

Bibliography

Häkkilä, J., Paananen, S., Suoheimo, M., & Mäkikalli, M. (2022). Pluriverse perspectives in designing for a cultural heritage context in the digital age. In Artistic Cartography and Design Explorations Towards the Pluriverse. Eds. Miettinen, S., et. al. Routledge, Taylor & Francis Group. pp. 134-143).

Moradi, F., Öhlund, L., Nordin, H., and Wiberg, M. (2020). Designing a Digital Archive for Indigenous People: Understanding the Double Sensitivity of Design. In Proceedings of the 11th Nordic Conference on Human-Computer Interaction: Shaping Experiences, Shaping Society (NordiCHI ’20), October 25–29, 2020, Tallinn, Estonia. ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3419249.3420174

Tuominen, I., Ballardini, R., Mähönen, J., & Pihlajarinne, T. (2023). Protecting and Accessing Indigenous Peoples’ Digital Cultural Heritage through Sustainable Governance and IPR Structures – the Case of Sámi Culture. Arctic Review on Law and Politics, 14, 194–219. https://doi.org/10.23865/arctic.v14.5809

Eetu Mäkelä, Thea Lindquist, Narges Azizifard, Julius Arnold
University of Helsinki, Finland

Exploring Co-Authorship Networks and Dynamics Among 17th Century Fruitbearing Society Members

Wednesday, 05/Mar/2025 4:40pm - 5:00pm
ID: 204 / Session SP 03: 3
Short paper (abstract) | 15-minute presentation with a 5-minute Q&A
Keywords: Temporal Network, Bibliographic Network, Co-authorship, Computational Bibliography, Early Modern Germany

In recent decades, numerous tools and metrics have been developed and widely applied in bibliometric and sociological research to describe the evolution of scientific disciplines and their social and cognitive dynamics. Unlike traditional qualitative methods employed in the sociology of science, these metrics rely on the quantitative analysis of factors such as authors, citations, keywords, journals, and other bibliometric indicators to extract the essence of contemporary science (Maltseva & Batagelj, 2022). One common approach is to analyze relationships between authors by examining their joint publications, which leads to the creation of co-authorship (co-occurrence) networks. These networks extend beyond sociological research and can be used to identify authorship communities and collaboration patterns across various disciplines, such as digital humanities (Yan & Ding, 2012; Rehbein, 2020; Laudel, 2002), which is the focus of our study. A co-authorship network is generated by directly linking actors based on their co-occurrence in the publications (Ahnert et al., 2021; Tiihonen et al., 2022).

Previous research by authors (Hill et al., 2019) has demonstrated the validity and potential of historical network analysis, such as that conducted using data from the English Short Title Catalogue (ESTC). The term "network" is often used informally by book historians—books in the seventeenth and eighteenth centuries were considered "networked technologies" (Greteman, 2021) because they were produced collaboratively, and those involved naturally formed networks with overlapping alliances, fostering both local and international connections. Examining these networks, at least qualitatively, has long been a central approach in the study of book history (Tiihonen et al., 2022).

Applying network analysis to early modern publications, by the 890 members of the Fruitbearing Society (Fruchtbringende Gesellschaft) (1617–1680), the first and largest cultural society in early modern Central Europe (Mäkelä et.al, 2024), is the main goal of our research. Co-authorship network analysis is highly informative for understanding the structure and patterns of collaboration among these members. By constructing both dynamic and time-specific co-authorship networks of these early modern publications, we can examine the factors influencing publication characteristics, such as genre, the number of collaborations per publication, and the scientific and political positions of the dedicatees. Additionally, we aim to determine if these networks were susceptible to historical transitions, especially during moments of political and social crisis (Ladd, 2021).

Gaining a deeper understanding of the publications associated with the Society's agenda is significant, as the Society had considerable influence on the development of early modern Central European thought and culture in the 17th century. The Society was committed to two main targets: cultivating virtue and promoting the development of the German language. It played a pivotal role in establishing German as a literary and scholarly language and served as a model for other 17th-century German language societies (Ball, 2010). In summary, we aim to explore how the collaboration patterns in member publications related to the Society's agenda evolved across its three major phases: the Köthen period (1617–1650), the Weimar period (1651–1662/67), and the Halle period (1667–1680). We will also investigate whether changes in the Society’s headquarters and leadership affected collaboration features such as members diversity, publication types, and differences across these three periods.

Bibliography

Maltseva, D. & Batagelj, V. (2022). Collaboration between authors in the field of social network analysis. Scientometrics 127, 3437–3470.

Yan, E. & Ding, Y. (2012). Scholarly network similarities: How bibliographic coupling networks, citation networks, cocitation networks, topical networks, coauthorship networks, and coword networks relate to each other. J Am Soc Inf Sci Tec, 63: 1313-1326.

Rehbein, M. (2020). Historical Network Research, Digital History, and Digital Humanities. In The Power of Networks (pp. 253-279). Routledge.

Laudel, G. (2002). What do we measure by co-authorships? Research Evaluation, 11(1), 3–15.

Ahnert R., Ahnert S.E., Coleman C.N. & Weingart S.B., (2021). The Network Turn: Changing Perspectives in the Humanities. Cambridge University Press.

Tiihonen, I. L. I., Ryan, Y. C., Pivovarova, L., Liimatta, A., Säily, T. & Tolonen, M. (2022). Distinguishing discourses: A data-driven analysis of works and publishing networks of the Scottish Enlightenment. In Digital Humanities in the Nordic and Baltic Countries Conference (pp. 120-134). CEUR-WS. org.

Hill, M. J., Vaara, V., Säily, T., Lahti, L. & Tolonen, M. (2019). Reconstructing intellectual networks: From the ESTC’s bibliographic metadata to historical material, in C. Navarretta, M. Agirrezabal, B. Maegaard (Eds.), Proceedings of the Digital Humanities in the Nordic Countries 4th Conference, volume 2364 of CEUR Workshop Proceedings, CEUR, Copenhagen, Denmark, 201–219.

Greteman, B. (2021). Making connections with Milton’s Epitaphium Damonis, in Making Milton, Oxford University Press, 31–41.

Mäkelä, E., Lindquist, T., Fard, N.A. & Arnold, J. (2024). Fruchtbringende Gesellschaft (1617–1680) Member Publication Patterns in the VD17. In Proceedings of the Digital Humanities in the Nordic and Baltic Countries Conference. DHNB Publications.

Ladd, J. R. (2021). Imaginative networks: Tracing connections among early modern bookdedications. Journal of Cultural Analytics, 6 (1).

Ball, G. (2010). Die Tugendliche Gesellschaft – zur Programmatik eines adeligen Frauennetzwerkes in der Frühen Neuzeit. In Sammeln, Lesen, Übersetzen als höfische Praxis der Frühen Neuzeit. Die Böhmische Bibliothek der Fürsten Eggenberg im Kontext der Fürsten und Fürstinnenbibliotheken der Zeit, edited by Jill Bepler and Helge Meise, 337–361. Wolfenbütteler Forschungen, vol. 126. Wiesbaden: Harassowitz.

Julia Matveeva[1], **Veli-Matti Pynttäri**[2], **Kati Launis**[2], **Leo Lahti**[1]
[1]University of Turku, Finland; [2]University of Eastern Finland, Finland

Subset Selection in Bibliographic Research: Exploring the Boundaries of Automated and Manual Curation

ID: 190 / Poster Session 2: 11
Poster and demo (abstract) with accompanying a 1-minute lightning talk
Keywords: Bibliographic metadata, Manual curation, Automated subset selection, Hybrid methodology, Finnish literature, 19th-century fiction.

The rise of computational methods in the humanities has opened new possibilities for large-scale analysis of bibliographic data, particularly in literary studies. [1,2,3,8] Combining quantitative analysis with qualitative interpretation often relies on the selection of subsets of items for a closer manual analysis.[7,9] However, effective subset selection, which is crucial for careful qualitative interpretation, remains a challenge. Traditionally, it has been based on manual inspection; more recently, automated approaches as a fast and scalable solution. Traditional manual curation ensures nuanced expert selection but can be very time-consuming and prone to subjective biases, while fully automated methods offer scalability but may overlook critical details or raise equivocal generalisations.[4,7] We explore the intersection of these complementary approaches by comparing automated and manual subset selection methods in the context of 19th-century Finnish fiction, using the National Finnish Bibliography, Fennica as a primary data source.[6] We evaluate the advantages, limitations, and biases inherent in each strategy. We hypothesize that a hybrid method—combining the strengths of both approaches—could provide an optimized solution for bibliographic research. We curated a list of first editions of Finnish fiction - or belles-lettres - published between 1809 and 1917, including novels, short stories, drama, poems, written in Finnish or Swedish. The manual curation process involved expert-driven analysis of bibliographic metadata. The manually curated list initially contained 4,468 titles, and it was further reduced to 2,788 titles after applying additional selection and exclusion criteria, and removing duplicates. The manual curation was particularly useful in reliably identifying first editions and filtering out irrelevant titles such as reprints or translations , that are not included in the intended subset. However, this process also relied on subjective decisions that could potentially obscure transparency and replicability, lead to biases, and have limited scalability. Conversely, the automated process yielded a list of 3,696 titles based on similar automated selection criteria. Enrichment with additional metadata, such as genre and Universal Decimal Classification codes, further expanded the list to 4,698 titles. Automated methods proved to be efficient and scalable, allowing for efficient updates and integration of new data sources. However, it also highlighted challenges in handling incomplete or inconsistent metadata, which led to the inclusion of unwarranted titles and the occasional misclassification of works that could be detected by additional manual curation. Our findings suggest that neither approach alone can fully address the complexities of bibliographic research, and we discuss the possibilities of combining the two; a hybrid approach, combining the strengths of both methods, emerges as the most effective strategy.

Bibliography

[1] Bode C. A World of Fiction. Digital Collections and the Future of Literary History. 2018, 37–58. University of Michigan Press.

[2] Lahti L. Open Data Science. Advances in Intelligent Data Analysis XVII. Lecture Notes in Computer Science 11191:31-39. Springer, India, 2018.

[3] Lahti, L., Marjanen, J., Roivainen, H., & Tolonen, M. (2019a). Bibliographic data science and the history of the book (c. 1500–1800). Cataloging & Classification Quarterly, 57(1): :5-23. Special issue. Routledge.

[4] Lahti L, Vaara V, Marjanen J, Tolonen M. (2019b). Best Practices in Bibliographic Data Science. In: Jantunen et al. (eds.). Proc. Research Data And Humanities (RDHUM) 2019 Conference: Data, Methods and Tools. Studia humaniora Ouluensia 17:57-65. University of Oulu.

[5] Mäkelä E, Lagus K, Lahti L, Säily T, Tolonen M, Hämäläinen M, Kaislaniemi S, Nevalainen T. (2020). Wrangling with Non-Standard Data. In: Reinsone S, Skadiņa I, Baklāne A, Daugavietis J (eds.), Proc. Digital Humanities in the Nordic Countries. CEUR Workshop Proceedings 2612:81-96, 2020.

[6] National Library of Finland (2024). Finnish national bibliography as open data. URL: http://data.nationallibrary.fi/bib/me/CFENNI (accessed 11 July 2024).

[7] Parente-Čapková, V., Launis, K., & Westerlund, J. (2023). Digitaaliset metadata-arkistot ja mesoanalyysi kulttuurisen vaihdon kartoittamisessa (Digital metadata archives and mesoanalysis in mapping cultural exchange). Kirjallisuudentutkimuksen aikakauslehti Avain, 20(1), 100–111.

[8] Tolonen, M., Lahti, L., Roivainen, H., & Ilomäki, N. (2016). Printing in a Periphery: a Quantitative Study of Finnish Knowledge Production, 1640-1828. 383-385. Abstract from Digital Humanities, Kraków, Poland.

[9] Tolonen, M., Lahti, L., Marjanen, J., & Roivainen, H. (2019). A Quantitative Approach to Book-Printing in Sweden and Finland, 1640-1828. Historical Methods, 52(1), 57-78. URL: https://doi.org/10.1080/01615440.2018.1526657

Mark Mets[1,2]
[1]Estonian Literary Museum, Estonia; [2]Tallinn University

Hands-on tutorial on GPT zero-shot annotation: example of Estonians’ relations to music survey

ID: 320 / WS04B: 2
Explorations of the dynamics of cultural phenomena in text corpora
Keywords: music-related attitudes, Estonian vernacular music practices, artificial intelligence in humanities, semantic and thematic analysis

This hands-on workshop will demonstrate the approach used in the presentation “Computational approaches with zero-shot annotating for survey analysis: case of Estonians’ relations to music”. We show how to use OpenAI API for prompt-based annotation of texts (but potentially also images), how to evaluate the results in comparison to human-annotations, and further discuss the efficacy of this approach and annotation-related problematics based on our study.

Mark Mets[1,2], Peter Dodds[3], Maximilian Schich[1]
[1]Tallinn University, Estonia; [2]Estonian Literary Museum, Estonia; [3]University of Vermont, USA

Attention Changes Towards Ukraine in 28 Languages on Twitter from 2008 to 2023

ID: 159 / **Poster Session 2: 20**
Poster and demo (abstract) with accompanying a 1-minute lightning talk
Keywords: Ukraine, Twitter, n-grams, attention, macroscope

The Russo-Ukrainian War has received significant academic interest, yet computational research using social media data, particularly with extensive temporal and linguistic coverage, remains limited. Our study addresses this gap by tracing the log-deviation from expected log-rank-frequency of "Ukraine" mentions on Twitter from 2008 to 2023, across 28 languages. We examine the patterns of attention before and after the 2014 and 2022 Russian invasions, revealing notable differences in the onset and decay of attention within and across languages. These variations highlight fundamental shifts in international support and regional sensitivities.

In February 2014, Russia's initial invasion of Ukraine triggered an ongoing conflict, which escalated dramatically in 2022 with a much larger-scale Russian invasion. Studying mentions of “Ukraine” on Twitter provides a useful proxy for the attention to the crisis. Our macroscopic approach leverages n-gram frequency data from Storywrangler.com, based on 10% of all tweets, focusing on the rank-frequency of “Ukraine” across multiple languages. This method provides a simple yet powerful lens for studying shifts in international perception.

Our analysis uncovers fundamental differences in the patterns of attention surrounding the 2014 and 2022 invasions. Not only was the global response to the 2022 invasion much more intense, but the sustained attention in European languages suggests deeper regional concerns. The decay of attention post-invasion shows stark differences between languages, with some continuing to focus on the crisis for a longer period. By focusing on log-deviation from expected frequency, we reveal specific, meaningful events and patterns over time that would otherwise remain obscured, including culturally and politically driven variations in attention across different linguistic communities.

In sum, our research offers a novel view of global attention to Ukraine on Twitter over 15 years. By examining language-specific responses to the 2014 and 2022 invasions, we provide valuable insights into the dynamics of international support and offer visual overview of the attention across languages and years. We demonstrate the pertinence of such macroscopic perspective for understanding global crises.

Mark Mets[1,2], **Taive Särg**[1], **Kadri Vider**[1], **Natali Ponetajev**[1], **Mari Väina**[1], **Kaarel Veskis**[1], **Mia Marta Ruus**[1], **Janika Oras**[1]
[1]Estonian Literary Museum, Estonia; [2]Tallinn University

Computational approaches with zero-shot annotating for survey analysis: case of Estonians' relations to music

ID: 314 / WS04A: 4
Explorations of the dynamics of cultural phenomena in text corpora
Keywords: music-related attitudes, Estonian vernacular music practices, artificial intelligence in humanities, semantic and thematic analysis

Since the 20th century, music research has increasingly focused on how people engage with music and how music affects them. The findings generally highlight the increasing role of recorded music, the positive influence of music on individuals' well-being as well as their physical and mental health, although some studies have also noted negative aspects of music (e.g., DeNora 2000; Turino 2008; Juslin & Sloboda 2011; Camlin et al. 2020; Agersnap et al. 2023).

In this presentation, we introduce our methodological approach to survey and essay analysis (cf. Karjus 2023) on how Estonians encounter music and their emotional attitudes toward it. We examine the texts submitted to the 2022 Estonian Folklore Archives' collection competition, "*Music in My Life*"*,* in which respondents shared their memories and experiences related to music. Approximately 300 senders responded to the competition in different ways, with 66 of responses submitted through our questionnaire platform Kratt where the answers were formatted question by question. The rest of the material was sent via e-mail or in handwritten form, allowing respondents to freely structure their answers. We have digitized the handwritten submissions and annotated the structure elements and metadata of free-form submissions.

We explore the thematic categories, activities related to music and various attitudes toward music, including those shaped by both direct exposure to music and indirect influences. For explorative data analysis as well as more specific categorization tasks we apply computational approaches, such as GPT-4o for machine-assisted zero-shot annotations that yielded statistically satisfactory results in filtering out music related sentences, identifying the topics and stances towards music (cf. Liyanage et al. 2023) .

Our preliminary findings from the group of younger respondents reveal that singing was the most frequently mentioned activity, listening was the predominant mode of engagement with music, and the phrase "I like" was most commonly used to express attitudes. These results gain deeper significance when compared to responses from other groups.

Bibliography

Agersnap, Anne; Lea Wierød Borčak, Thomas Husted Kirkegaard, Katrine Frøkjær Baunvig 2023. Danish Communal Singing Culture in the 2020ies: A Survey of Singing Habits among Adult Danes in Denmark. Grundtvig eStudies Pamphlets, No. 2.

Camlin, David A.; Helena Daffern & Katherine Zeserson 2020. Group Singing as a Resource for the Development of a Healthy Public: A Study of Adult Group Singing. Humanities and Social Sciences Communications 7: 60.

DeNora, Tia 2000. Music in Everyday Life. Cambridge: Cambridge University Press.

Juslin, Patrik N. & John Sloboda (eds.) 2011. Handbook of Music and Emotion. Oxford: Oxford University Press.

Karjus, Andres 2023. Machine-assisted mixed methods: augmenting humanities and social sciences with artificial intelligence. arXiv preprint arXiv:2309.14379.

Liyanage, Chandreen; Ravi Gokani & Vijay Mago 2023. GPT-4 as a Twitter data annotator: Unraveling its performance on a stance classification task. Authorea Preprints.

Turino, Thomas 2008. Music as Social Life: The Politics of Participation. Chicago: University of Chicago Press.

Mark Mets[1,2], **Kadri Vider**[1], **Taive Särg**[1], **Katrine Frøkjær Baunvig**[3], **Mari Väina**[1]
[1]Estonian Literature Museum; [2]Tallinn University; [3]Aarhus University

Explorations of the dynamics of cultural phenomena in text corpora

ID: 260
Full-day conference-themed workshop
Keywords: computational methods, cultural analytics, macroscopic, text mining, NLP

Texts, whether written, spoken, or visual, are dynamic records of cultural expression that both reflect and shape the communities from which they emerge. Through careful analysis, corpora can thus illuminate the intricate interplay between culture, language, and time, offering a rich resource for understanding cultural change and continuity. Computational approaches to large collections of texts allow researchers to trace how cultural phenomena emerge, distribute, spread, and transform within a society. A cultural phenomenon here refers to any concept, behavior, or experience that is shaped by the beliefs, values, and practices of a particular society or group. Importantly, computational approaches enable the detection of patterns at a scale that may be invisible by close reading only. The analysis of large-scale text corpora may reveal cultural processes that are otherwise difficult to observe, the diachronic change or synchronic variability of the models of communication, action, and values in various texts of various origins.

The workshop invites to discuss the experiences of studying diverse corpora, whether separately or in combination. We aim to gather the experience across countries, approaches and datasets to share the knowledge of large-scale textual analysis with an emphasis on discovering underlying cultural patterns, different values, beliefs and practices and different ways of modeling the world. It includes taking into consideration and comparing the variety in cultural patterns one may observe at a scale, the diversity of groups being observed, methods, and possible biases.

The analysis of cultural phenomena involves both a diachronic dimension capturing the changes across time, and a synchronic approach that allows us to detect variation of features across or within the corpora. Diachronic analysis may reveal temporal shifts or changes in expressions, discourse, mindsets and societal values. Synchronic analysis affords comparison of such expressions, discourses, mindsets and values between the text collections or their subsets. What differs across text corpora and cultures they represent are both the form and the underlying meanings. It is relevant to bring out the differences in form, such as linguistic differences in word forms, whilst our focus is on grasping these features as proxy for more substantial differences.

This workshop invites participants to explore the diversity of textual sources, from vernacular to elite, to identify different understandings in culture and to find commonalities where possible. Each text corpus reflects distinct social discourses, offering insights into how communities and historical periods conceptualize and negotiate cultural meanings. Texts, as cultural artifacts, actively shape meaning, embedding the power dynamics of social groups, institutions, and ideologies. By examining a range of genres and media, researchers can track shifts in cultural discourse in response to political, technological, and social changes, revealing how diverse communities contribute to the evolution of cultural meanings. Studying cultural patterns at scale may allow us to explore the emergence and adoption of global trends as well as the dissemination of ideas among specific groups. Yet, every corpus has specific people or groups behind it, and we emphasize the relevance of recognizing whose voices the texts may represent. Moreover, rarely can we have ideal datasets for computationally understanding culture. The methods for examining cultural dynamics at scale must consider these limitations but can also provide ways to recognise and mitigate these shortcomings.

The methods to grasp the texts at large put an emphasis on exploratory data science, by first looking at the world and being open to possibilities of what we might find. This does not exclude hypothesis or theory driven approaches, but rather combines them to rediscover large scale changes that were hard to grasp without the help of computers. It may be thought of as a back-and-forth interaction between distant and close reading, gaining much from the diversity offered by interdisciplinarity. The workshop discusses this in the context of visualization of large textual corpora as systems of relations, whether as networks, embedding spaces, trends or by other means. This includes text enrichment through sentiment or stance detection, topic modelling that may reveal hidden (embedded) meanings and their change. Classical NLP methods of textual comparison, LLMs and now generative models have paved the way for more and more approachable and versatile methods of analysis. Prompt based exploration of texts provides a new level of interaction where one is capable of asking about the corpora questions in an intuitive manner. Last but not least, it is a challenge to do it in a statistically rigorous way.

This workshop is aimed at researchers who are engaged in or are interested in learning more about computational studies of text corpora to better understand underlying cultural dynamics. As an outcome of the workshop we hope to get an overview of the variety of approaches currently in use among the researchers for exploring the content aspects and their variation within large corpora. We also look forward to meeting researchers with similar interests, gaining inspiration, and fostering possible future collaboration.

Giedrė Milerytė-Japertienė
National museum of Lithuania, Lithuania

The Lithuanian National Museum: Advancing Digital Heritage Initiatives

ID: 305 / **WS02: 2**
Innovations and New Interactions through Digital Cultural Heritage in GLAM Sector
Keywords: National museum, collaboration iniciatives, digitization, virtual tures

The Lithuanian National Museum (LNM), home to the largest collection of cultural heritage artifacts in the country, comprising over 1.3 million items, was a relatively late entrant into the digitization process. It joined the national Lithuanian Integral Museum Information System (LIMIS) only in 2020. Over four years of digitization efforts, more than 23,000 cultural assets have been made accessible to users. Despite its delayed start, the museum is making significant strides in this area, aiming to catch up by both increasing digitization output and ensuring high-quality data presentation.

One key advantage of this later entry is the ability to avoid revisiting metadata issues, such as license assignments, that earlier adopters faced. Additionally, LNM has implemented a curated approach to the digitization process, focusing on thematic or logically connected groups of objects rather than digitizing them sequentially. This strategy ensures that digital collections are coherent and meaningful to users.

By leveraging the latest technologies, the museum is creating curated digital content for exhibitions and public engagement. Digitized objects are integral to virtual tours, online exhibitions, and even physical displays where 3D-printed models serve as replicas of authentic artifacts. Moreover, the museum collaborates with companies specializing in virtual and augmented reality game development, incorporating its cultural assets into interactive experiences.

The presentation will explore LNM's innovative initiatives, highlighting specific examples of how collaboration with academic and business sectors empowers the public to engage with and explore cultural heritage through its digital representations. These efforts demonstrate the museum’s commitment to modernizing heritage preservation and making it accessible in creative and meaningful ways.

Daniele Monticelli[1], Krister Kruusmaa[1,2]
[1]Tallinn University, Estonia; [2]National Library of Estonia

Visualizing Literary History: From Bibliographic Data to Network

Wednesday, 05/Mar/2025 5:00pm - 5:20pm
ID: 231 / Session SP 03: 4
Short paper (abstract) | 15-minute presentation with a 5-minute Q&A
Keywords: network analysis, translation studies, library data

The paper aims to address the methodological and empirical possibilities, as well as challenges of visualizing historical processes from bibliographic metadata in library catalogues. Our discussion is based on a case study, the interactive Network of Estonian Translated Literature (NETL) which was created as a cooperation between the Digilab of the National Library of Estonia and Tallinn University's Research Grant "Translation in History, Estonia 1850-2010: Texts, Institutions, Agents and Practices".

While network theory (Folaron & Buzelin 2007; Paloposki 2018) and the visualization of translation spaces (Hofeneder 2022, Zhai et al. 2020, Tanasescu 2020), have recently gained the attention of TS scholars, NETL stands out both due to its scale and completeness. This network is based on the Estonian National Bibliography which encompasses comprehensive and well-coded information on all known works of literature in the Estonian language. Extensive metadata curation at the Digilab (Kruusmaa et al. 2025) has enabled the visualization of historical connections between authors and all known Estonian literary translators within a single, dense network. Consisting of more than 12 000 nodes from the past two centuries, NETL can be searched and navigated as well as filtered by language, genre and temporally.

In our paper, we will discuss NETL as an innovative approach to a data-driven study of historical issues, situated at the intersection of digital humanities, translation studies and evolutionary cultural analysis. Collaboration between traditional humanities researchers (historians, literary, and translation scholars) and those in data analytics and visualization has allowed us to bridge key questions in historical translation studies with computational methods. We will argue that the network reveals a range of possibilities and challenges in data-driven research on translation history (Wakabayashi 2019).

Our study applies network analysis and interactive visualization to explore the evolution of translation history in Estonia from 1900-2019, demonstrating how this approach effectively combines macro-level computational analyses with meso- and micro-level qualitative insights. This perspective reveals shifts in literary exchange shaped by historical transformations. The development of translation networks reflects broader cultural and political transitions, with periods of restriction and expansion influencing the circulation of translated works. While state policies have shaped translation flows, the persistence of certain literary connections suggests deeper continuities that transcend political shifts. The analysis highlights how translation networks not only reflect historical disruptions but also reveal patterns of resilience and adaptation, offering a nuanced perspective on cultural and literary evolution.

Finally, we will argue that NETL is both an instrument for research and a good example of knowledge transfer between the academia, memory institutions and the general public, as the network is conceived as an open resource and user-friendly interface for all the users interested in foreign literature, translation and translators, cultural networks and Estonian cultural history.

Bibliography

Current NETL prototype: https://data.digar.ee/netl

Folaron, Deborah et Hélène Buzelin 2007. "Introduction: Connecting Translation and Network Studies". Meta 52(4), 605–642.

Hofeneder, Philipp. 2022. "A cartography of translation. Visualizing translation spaces", Translation Spaces 11(2), 157-183.

Kruusmaa, Krister, Tinits, Peeter, Nemvalts, Laura (2025). "Curated Bibliographic Data: the Case of the Estonian National Bibliography". Journal of Open Humanities Data, 11

Lahti, Leo, Marjanen, Jani, Roivainen, Hege, & Tolonen, Mikko (2019). "Bibliographic Data Science and the History of the Book (c. 1500–1800)." Cataloging & Classification Quarterly, 57(1), 5–23.

Paloposki, Outi 2018. "The Missing Needle: Bibliographies, Translation Flows and Retranslation", in: Translating Scandinavia. Scandinavian Literature in Italian and German Translation, 1918-1945, edited by Bruno Berni, Anna Wegener. Roma: Edizioni Quazar, 15-28.

Poupaud, Sandra, Pym, Anthony, Simón, Ester Torres (2009), "Finding translations. On the use of bibliographical databases in translation history", Meta, 54(2): 264-278.

Tanasescu, Raluca. (2020). "Chaos out of Order: Translations of American and Canadian Contemporary Poetry into Romanian before 1989 from a Complexity Perspective". Chronotopos, 2, 64-94. Article 3. https://doi.org/10.25365/cts-2019-1-2-5

Wakabayashi, Judy 2019. "Digital approaches to translation history". Translation & Interpreting 11(2), 132-145.

Alex Zhai, Zheng Zhang, Amel Fraisse, Ronald Jenn, Shelley Fisher Fishkin, and Pierre Zweigenbaum 2020. "TL-Explorer: A Digital Humanities Tool for Mapping and Analyzing Translated Literature". Proceedings of the 4th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature, edited by Stefania DeGaetano, Anna Kazantseva, Nils Reiter, Stan Szpakowicz: 161-171. https://aclanthology.org/2020.latechclfl-1.20/.

Filippo Mosca[1], **Jakub Gomułka**[2]
[1]University of Rome Tor Vergata, Italy; [2]University of Kraków, Poland

AI-Assisted Philosophy: How Wittgenstein Scholars Engage with Customised GPT Models

Friday, 07/Mar/2025 2:10pm - 2:30pm
ID: 149 / Session LP/SP 09: 3
Short paper (abstract) | 15-minute presentation with a 5-minute Q&A
Keywords: Philosophy, Wittgenstein, LLM Technology, Customised GPT Models

In our presentation, we will discuss the potential uses of LLM technology in the field of philosophical inquiry.

The subject of our research is a range of customised versions of OpenAI's ChatGPT, which we prepared in early 2024. These customised GPT models are specifically designed to use the original writings of the philosopher Ludwig Wittgenstein as the basis for their responses to user questions or prompts, potentially allowing for an interactive exploration of his philosophy.

The purpose of this interactive exploration is twofold: to lay the groundwork for an AI-based chatbot that can assist users in studying and researching Wittgenstein's writings; to gain some insights into AI's ability to engage in philosophical dialogue and interact with complex philosophical texts.

Tractarian Wittgenstein is a GPT model based on the German text of Wittgenstein's Tractatus logico-philosophicus. Wittgensteinian Oracle, on the other hand, is a series of GPT models, each based on a particular manuscript or typescript from Wittgenstein's Nachlass (included in the 5000 pages of the Discovery Project).

The aim of this research is to gain insight into how Wittgenstein scholars interact with our customised models. To this end, we have collected free reports from Wittgenstein scholars who have been invited to test the models according to their own preferences and interests. As this is only the initial stage of our project, the feedback gathered so far is still limited. However, certain trends can already be observed.

One of these trends is the users' cautious and skeptical approach towards the models, which begins to take shape as initial interactions clearly show that they cannot be treated as 100% reliable authorities. In particular, it has been observed that the models often attribute theses to Wittgenstein that he does not actually support. This is partly due to their inability to recognize his use of sarcasm and their failure to consider the broader context behind individual remarks or isolated groups of remarks. Nevertheless, we noted a shift towards a more positive assessment of the technology's capabilities as users gain more experience with it. An element of surprise has emerged: experts reported that although the models' responses were not free of inaccuracies and oversimplifications, some responses seemed interesting and even to some extent potentially useful for their research practice. On the one hand, in fact, errors are sometimes seen as valuable, because they invite the scholar (or the student, in educational settings) to engage creatively in correcting them, which may lead to insights that might be missed or might be more difficult to reach through conventional reading. For instance, one scholar noted that correcting an unclear and unsatisfactory answer from the model on the complex relationship between the concepts of "world" and "reality" in the *Tractatus* led him to reconsider the question from an unusual perspective. On the other hand, errors are seen as tolerable because they are compensated by some remarkable capabilities.

In addition to technical features such as response speed, consistency and memory within a single session, and the ability to provide real-time translation of information into multiple languages, scholars seem to particularly value two key aspects of these models. The first aspect is their role as advanced research tools. The models prove especially valuable in facilitating linguistic analysis. One scholar in particular, who has long dealt with the evolution of Wittgenstein's style, was impressed by the model's ability to analyse the stylistic properties of Wittgenstein's Nachlass documents. The models also proved very useful for navigating philosophical topics and summarising complex ideas. These functions help users to organise their thoughts more efficiently and to access Wittgenstein's philosophy in a structured (although not always precise and reliable) way. The second aspect is their function as creativity stimulators. On the one hand, their very nature as interactive systems encourages the user to think “dialogically”. For example, one scholar noted how repeated interactions with these models improved their ability to formulate questions, which can be beneficial not only in the context of individual GPT sessions, but also in broader research settings. On the other hand, they stimulate new insights by suggesting unexpected and creative connections.

Another notable trend that emerges from the reports is the curiosity about the cognitive limits of the model. Wittgenstein scholars, especially in more advanced stages of interaction, seem interested not only in how the models process and deliver philosophical content, but also in whether (and to what extent) they can handle ambiguity, abstract reasoning, contextual meaning, and whether (and to what extent) they can reflect on their own reasoning processes and limitations. For example, some scholars have found it fruitful to ask the model how it would prefer to be questioned. It might be therefore expected that the potential repeated use of these systems by experts in philosophy could indirectly contribute to the ongoing debate about the future development of LLMs, where philosophical questions about cognition and metacognition intersect with engineering challenges.

Bibliography

Azaria, A. (2022). ChatGPT Usage and Limitations. (https://hal.science/ hal-03913837)

Ghosh, L. (2023). Will Google Gemini outdo GPT-4? WIRE19. (https://wire19.com/ google-gemini-vs-gpt-4)

Gangopadhyay, N., Grève, S. S., & Pichler, A. (2023). “A Complex Philosophical Oeuvre and Its Complex User Community: Reflections on the Past, Present, and Future Digitisation of Wittgenstein’s Philosophical Writings”. Digital Humanities in the Nordic and Baltic Countries Publications, 5(1): 105-120.

Hu, K. (2023). ChatGPT Sets Record for Fastest-Growing User Base—Analyst Note. (https://www.reuters.com/technology/chatgpt-sets- record-fastest- growing-user-base-analyst-note-2023-02-01)

Mosca, F. & Gomułka, J. (2024), “What Wittgensteinian GPTs Can’t Do”, in Contributions of the 45th International Wittgenstein Symposium, edited by Yannic Kappes, Asya Passinsky, Julio De Rizzo, Benjamin Schnieder, Österreichische Ludwig

Wittgenstein Gesellschaft, Kirchberg am Wechsel (ISSN 1022-3398).

Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). Improving language understanding by generative pre-training.

Schwitzgebel, E., Schwitzgebel, D., & Strasser, A. (2023). Creating a Large Language Model of a Philosopher. ArXiv, abs/2302.01339.

Truelove, K (2023). On the Nature of Artificial Things. TrueSciPhi.AI (https://www.truesciphi.ai/p/on-the-nature-of-artificial-things)

Weinberg, K. (2023). Philosophical Uses for LLMs: Modeling Philosophers, Daily Nous, (https://dailynous.com/2023/12/05/philosophical-uses-for-llms-modeling-philosophers)

Wittgenstein, L. (2000), Wittgenstein's Nachlass: The Bergen Electronic Edition, edited by Claus Huitfeldt, Oxford, Oxford University Press.

Nasrin Mostofian, Anna Foka
Uppsala University, Sweden

The use and usefulness of AI in CHIs

Wednesday, 05/Mar/2025 4:40pm - 5:00pm
ID: 129 / Session SP 01: 3
Short paper (abstract) | 15-minute presentation with a 5-minute Q&A
Keywords: Cultural Heritage Institutions, Classification, AI models, Bias, Diversity

According to Heritage Fund UK, “Heritage organisations are using Artificial Intelligence (AI) in three key areas of their work: heritage and collections management, use and research; visitor experience; and, business operations and management” (Heritage Fund, n.d.). From reconstructing and restoring masterpieces and paintings, to detecting archaeological sites, AI has proven to be beneficial in CHIs (Europarl briefing, 2023). The opportunities and challenges that the use of AI and Machine Learning (ML) creates for the curation and classification of collections in Cultural Heritage Institutions (CHIs) has been the topic of interest for various research studies in the past years (see for instance Tzouganatou, 2018; Murphy & Villaespesa, 2020; Foka et al. 2022; Foka et al. 2023). Therefore, aforementioned research shows that from a technical point of view, there is good potential for the application of pioneering technologies such as AI and ML in the heritage sector. On the other hand, from a social, cultural, and political perspective, there is a considerable risk of transferring existing bias when these technologies are applied; therefore leading to social inequalities. Significant work remains to establish a symbiotic relationship between AI techniques and CHIs. Ideally, this symbiosis will enable the harnessing of these technologies to facilitate and promote the curation of cultural heritage collections, aligning with the values of an increasingly diverse, equitable, and inclusive global society.

This paper’s objective is to address the issue of bias within CHIs and to explore the potential applications of AI across various sectors of CHIs for the classification and curation of collections. We focus on diversity and inclusion from the lens of gender and ethnicity. Given the complexity of diversity issues in CHIs and the nascent state of AI usage in these institutions, a theoretical framework is needed to guide the application of AI in CHIs.

The current practices of AI in various CHIs worldwide provide strong evidence that the future of CHIs will benefit from this technology in multiple ways, not least in supporting a faster curation of collections by automating processes such as classification, categorisation, metadata etc. In this paper we begin by introducing emerging technologies, such as AI models, as not fully understood into the task of classification. We argue that this process introduces complexity which is further compounded when, in the context of power dynamics; the question of who determines the criteria for classification intersects with the question of who decides what data are used to train these AI models. Finally, we advocate that it is imperative to develop a theoretical and empirical framework for the application of AI in meaningful ways that guide CHIs towards an equitable classification system and promote diversity as inclusion, both now and in the future. Our methodology draws on relevant cultural theories such as feminist studies, decolonisation, intersectionality and critical cultural theories, and suggests a corrective approach for the implementation of AI-based tools in CHIs based on these theories. Cultural theories like feminist studies and decolonization are increasingly integrated into AI and heritage collection research. Scholars argue that bias is inherent in cultural heritage collections and can be amplified by AI pipelines. Researchers are exploring how to critically apply AI to label and interpret archival material, aiming to make cultural heritage more inclusive and representative of diverse perspectives. This involves addressing issues of digital cultural colonialism, gender bias, and power structures in heritage discourses. The goal is to implement bias mitigation techniques throughout the process, from collection to curation, to support meaningful curation, embrace diversity, and cater to future heritage audiences (see for instance Smyth et al., 2021; Foka & Griffin, 2024).

In addition to studying pertinent theories and applying them in our research, we refer to some finds from semi-structured interviews with curators in CHIs across Sweden to gain a deeper understanding of the status of AI in this field. The results of these interviews reveal that in spite of the will for the application of AI in CHIs in practice AI and ML are not being practiced in these institutions. During the courses of interviews most interlocutors express their will for using AI on a personal level, and mention factors such as the lack of clear guidelines that clarify dos and don’ts, education, and budget as some of the obstacles to using AI at their institutions.

To conclude, we propose an approach towards the productive implication of AI in CHIs and suggest, among other things, a Human-in-the loop approach. This approach relies on the fact that AI applications are not inherently destined to perpetuate existing bias. Rather, as a product of human intelligence, they reflect the data they are trained on. We argue that timely intervention and the involvement of a diverse human approach in supervising AI techniques can potentially enrich collections in CHIs and raise awareness of inclusion issues that have been overlooked for decades.

Bibliography

Europarl briefing (2023). Artificial intelligence in the context of cultural heritage and museums: Complex challenges and new opportunities. https://www.europarl.europa.eu/RegData/etudes/BRIE/2023/747120/EPRS_BRI(2023)747120_EN.pdf

Foka, A., Attemark, J., & Wahlberg, F. (2022). Women’s Metadata, Semantic Web, Ontologies and AI: Potentials in Critically Enriching Carl Sahlin’s Industrial History Collection. In Emerging Technologies and Museums (1st ed., pp. 65-). Berghahn Books. https://doi.org/10.3167/978180073374900

Foka, A., Eklund, L., Løvlie, A. S., & Griffin, G. (2023). Critically assessing AI/ML for cultural heritage: potentials and challenges. In S. Lindgren (Ed.), Handbook of Critical Studies of Artificial Intelligence (pp. 815-825). Edward Elgar Publishing. https://doi.org/10.4337/9781803928562.00082

Foka, A., & Griffin, G. (2024). AI, Cultural Heritage, and Bias: Some Key Queries That Arise from the Use of GenAI. Heritage, 7(11), 6125–6136. https://doi.org/10.3390/heritage7110287

Heritage Fund. (n.d.) Artificial Intelligence: a digital heritage leadership briefing.

https://www.heritagefund.org.uk/about/insight/research/artificial-intelligence-digital-heritage-leadership-briefing

Murphy, O., & Villaespesa, E. (2020). AI: A museum planning toolkit. Goldsmiths, University of London.

Smyth, H., Nyhan, J., Flinn, A., Dunn, S., & Schuster, K. (2021). Opening the ‘Black Box’ of Digital Cultural Heritage Processes: Feminist digital humanities and critical heritage studies. In Routledge International Handbook of Research Methods in Digital Humanities (1st ed., Vol. 1, pp. 295–308). Routledge. https://doi.org/10.4324/9780429777028-22

Tzouganatou, A. (2018). ’Can heritage bots thrive? Toward future engagement in cultural heritage’. Advances in Archaeological Practice, 6(4), Cambridge: Cambridge University Press, 377-383.

Mahdi Munshi, Vaibhav Agarwal, Viivi Pentikäinen, Krista H Lagus
University of Helsinki, Finland

Applying Self-Organizing Maps (SOM) to PERMA-H Framework: Analysing Well-Being Expressions in Suomi24 Online Discussions

Friday, 07/Mar/2025 1:50pm - 2:10pm
ID: 258 / Session LP/SP 09: 2
Short paper (abstract) | 15-minute presentation with a 5-minute Q&A
Keywords: Self-Organazing Maps, Well-Being Theory, PERMA-H

Introduction. This paper explores the application of the Self-Organizing Map (SOM) algorithm to analyse large-scale textual data through the lens of the PERMA-H well-being framework. SOM, an unsupervised machine learning method known for clustering and visualizing high-dimensional data, offers a powerful approach to uncovering latent structures in unstructured text corpora (Kohonen et al., 2000; Lagus et al., 2004). In this study, we seek to understand "**how do the six components of well-being manifest in forum discussions, and what patterns emerge when these components are clustered using SOM**". For this purpose, we apply SOM to user-generated content from Suomi24, one of Finland's largest online discussion platforms, with posts dating back to 2001 (Lagus et al., 2016). Our analysis focuses on how well-being is expressed in Finnish language forums using the multidimensional PERMA-H model, which includes Positive Emotion, Engagement, Relationships, Meaning, Accomplishment (cf. e.g. Donaldson et al., 2022) and a later addition, Health (Norrish et al., 2013).

Building on the work of Honkela et al. (2014), we extend sentiment analysis beyond traditional one-dimensional approaches by integrating a hybrid methodology that combines SOM with the PERMA-H framework. Our approach allows for identifying complex emotional and psychological patterns in written texts while also addressing the linguistic challenges of the Finnish language. Through this novel combination of methodologies, the study reveals how well-being is articulated, shared, and transformed in public discourse over time, offering new insights into the role of online communities in reflecting societal well-being.

Corpus. To achieve this, we obtained the Suomi24 online discussion forum corpus from Kielipankki (Suomi24 corpus 2001–2020). The corpus contains 21 subforums on the main level, including Suhteet (Relationships), Talous (Economy), and Ajanviete (Leisure). It divides hierarchically into further levels of subforums and eventually discussion threads, where posts are followed by comments to them. Along with texts of the posts, the corpus provides metadata such as authorship (anonymous or registered username), post date and time, complemented with linguistic features like lemmatized words, part-of-speech tagging, and sentiment analysis.

Methodology. The overall approach outlined here can be characterized as a *theory-driven dictionary-based text analysis method*. We first defined a PERMA-H dictionary for Finnish as follows: For each of the 6 PERMA dimensions, ChatGPT was carefully prompted to provide a hundred Finnish words that express the meaning of that PERMA-H component. Next a Finnish PERMA expert filtered and refined these lists, arriving at 100 lemmas that describe each of the wellbeing components. These 6x100 words make up of our PERMA-H dictionary.

Data encoding. In applying the PERMA-H we decided to treat a subforum's discussion in a particular year as "one document" from that subforum, resulting in 420 documents (20 years x 21 subforums). Next, for each subforum and for each PERMA-H component we calculated a score representing how prevalent that wellbeing quality was for that subforum. This was calculated by counting the number of times that a word (lemma) belonging to the dictionary entry for that component was encountered in the subforum, divided by the total number of word *tokens* within that same subforum. The resulting number we call "perma indicator" for that component, for that subforum.

SOM. Next, these six indicators per each subforum were given as input for the Self-Organizing Map (SOM) analysis. The SOM algorithm was applied using the "kohonen" package in R, with a hexagonal grid of 5x4 neurons. The SOM was trained for over 25,000 iterations with a decaying learning rate (from 0.05 to 0.01), and with default parameters for the neighbourhood function. The resulting SOM can now be examined to find out both general trends and specific anomalies in wellbeing expression across different subforums and time periods.

To examine possible clustering structure among the subforums, we used the U-Matrix visualization, reflecting how the wellbeing expression changes when moving from a map region to a nearby region. Component planes were also generated to visualize the how individual PERMA-H component changes across the SOM grid, helping us identify which wellbeing dimensions were prominent in certain forums.

Results and analysis. The results of the preliminary analysis reveal dependencies between the PERMA-H wellbeing dimensions, that characterize the subforums. For example, the forum *Ajanviete* (Leisure) consistently scored high in positive emotions, as well as exhibited high engagement levels, shedding light on the importance of leisure activities on one's wellbeing.

Similarly, *Matkailu* (Travel) was prominent across multiple dimensions, particularly positive emotion and meaning, suggesting that travel-related discussions provide users with happiness and a sense of purpose. This intersectionality was evident in the component planes of the SOM, indicating the multi-faceted nature of wellbeing. The forum *Talous* (Economy) also provided insights into the importance of public discourse in reflecting the wellbeing of societies at large. The fluctuation in the positive emotion and accomplishment scores within *Talous* (Economy) aligns with real-world economic events, such as the decline leading up to the 2008 financial crisis, demonstrating how digital forums mirror public sentiment and their importance in gauging societal challenges early and ensuring collective wellbeing.

Discussion. A key takeaway of this experiment is that the Self-Organizing Map (SOM) proves to be an effective model for interpreting multidimensional data, such as wellbeing components within forum discussions. By organizing complex, high-dimensional data into a visual grid, SOM allows for clear identification of patterns that can help in uncovering trends and relationships that may not be immediately visible through traditional analysis.

We consider this work as a step forward in understanding 'wellbeing informatics', a concept which was suggested in (Lagus et al., 2012) where the SOM was utilized to obtain a visualization of wellbeing states from a heterogeneous set of wellbeing-related measurements.

The development of the PERMA-H dictionary as well as its operationalization to measuring discussions opens a way towards further applications and wellbeing analyses from different corpora, hopefully leading to a more nuanced and multifaceted understanding of qualities of conversations as well as their evolution over time. This research provides substantial implications for future studies in social science, text mining, and wellbeing theory.

Bibliography

Donaldson, S. I., van Zyl, L. E., & Donaldson, S. I. (2022). PERMA+ 4: A framework for work-related wellbeing, performance and positive organizational psychology 2.0. Frontiers in psychology, 12, 817244. https://doi.org/10.3389/fpsyg.2021.817244.

Honkela, T., Korhonen, J., Lagus, K., & Saarinen, E. (2014). Five-dimensional sentiment analysis of corpora, documents and words. In Advances in Self-Organizing Maps and Learning Vector Quantization: Proceedings of the 10th International Workshop, WSOM 2014 (pp. 209-218). Springer International Publishing. https://doi.org/10.1007/978-3-319-07695-9_20.

Kohonen, T., Kaski, S., Lagus, K., Salojarvi, J., Honkela, J., Paatero, V., & Saarela, A. (2000). Self-organization of a massive document collection. IEEE Transactions on Neural Networks, 11(3), 574-585. https://doi.org/10.1109/72.846729.

Lagus, K., Kaski, S., & Kohonen, T. (2004). Mining massive document collections by the WEBSOM method. Information Sciences, 163(1-3), 135-156. https://doi.org/10.1016/j.ins.2003.03.017.

Lagus, K., Ruckenstein, M. S., Pantzar, M., & Ylisiurua, M. J. (2016). Suomi24: muodonantoa aineistolle. (Suomi24 — Giving Shape to Suomi24). (Valtiotieteellisen tiedekunnan julkaisuja; No. 10). Helsingin yliopisto. http://hdl.handle.net/10138/163190.

Lagus, K., Vatanen, T., Kettunen, O., Heikkilä, A., Heikkilä, M., Pantzar, M., & Honkela, T. (2012). Paths of well-being on self-organizing maps. In Advances in Self-Organizing Maps: 9th International Workshop, WSOM 2012 (pp. 345-352). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-35230-0_35.

Norrish, J. M., Williams, P., O'Connor, M., & Robinson, J. (2013). An applied framework for Positive Education. International Journal of Well-being, 3(2), 147-161. https://doi.org/10.5502/ijw.v3i2.2.

Suomi24 corpus 2001–2020 (2021) [Dataset]. Kielipankki – The Language Bank of Finland. https://www.kielipankki.fi/corpora/suomi24/.

Ghulam Mustafa
International Islamic University Islamabd, Pakistan

Use of Artificial Intelligence (AI) in Open and Distance Learning (ODL) institutions: Opportunities, Challenges, and the Way Forward

ID: 279
Workshop on Digital Humanities and Social Sciences/Cultural Heritage (DHSS/DHCH) in Higher Education
Keywords: AI, Open and Distance Learning, opportunities, challenges, strategies

This paper examines the use of Artificial Intelligence (AI) in Open and Distance Learning (ODL) institutions, discussing the opportunities and challenges it presents, as well as strategies for effective implementation in the future. AI has the potential to offer personalized learning experiences, increase engagement, and streamline administrative processes in education. However, it also brings concerns related to data privacy, bias, and the need for substantial infrastructure investments. The purposive sampling technique selected 5 faculty members from different universities. A semi-structured interview guide was developed to get data from the participants. Data was analyzed thematically by facilitation of NVivo 14. It aims to provide a comprehensive overview of AI's current state in ODL, explore the implications of its adoption, and suggest actionable steps for institutions to address these challenges. The study concluded. AI can provide and enhance academic and management skills of open distance learning.

Bibliography

Chen, X., & Wang, X. (2020). "AI in Open and Distance Learning: Opportunities and Challenges." Journal of Educational Technology & Society, 23(3), 1-11.

Holmes, W., & Tuomi, I. (2021). "The Ethics of AI in Education: Practices, Challenges, and Prospects." AI & Society, 36, 543-554.

Popenici, S. A. D., & Kerr, S. (2017). "Exploring the Impact of Artificial Intelligence on Teaching and Learning in Higher Education." Research and Practice in Technology Enhanced Learning, 12(1), 1-13.

Selwyn, N. (2019). "Should Robots Replace Teachers? AI and the Future of Education." Digital Education Review, 35, 1-17.
Zawacki-Richter, O., Marín, V. I., Bond, M., & Gouverneur, F. (2019). "Systematic Review of Research on Artificial Intelligence Applications in Higher Education – Where Are the Educators?" International Journal of Educational Technology in Higher Education, 16(1), 39.

Alamri, A. (2021). The Impact of AI in Enhancing Accessibility in Education. International Journal of Educational Technology, 7(2), 45-59.

Baggaley, J. (2013). Distance Education in Developing Countries. The Handbook of Distance Education, 3rd edition, 115-127.

Bates, A. W. (2015). Teaching in a Digital Age: Guidelines for Designing Teaching and Learning. BCcampus.

Chen, G., Cheng, W., & Chen, R. (2020). Personalized Learning in Online Courses: The Role of AI. Journal of Educational Computing Research, 58(6), 1103-1121.

Eynon, R., & Malmberg, L.-E. (2020). Understanding Learning and Learning Design in MOOCs: A Measurement-Based Approach. The Internet and Higher Education, 44, 100724.

Floridi, L., et al. (2018). AI4People: An Ethical Framework for a Good AI Society. Minds and Machines, 28(4), 689-707.
Laurillard, D. (2012). Teaching as a Design Science: Building Pedagogical Patterns for Learning and Technology. Routledge.
Luckin, R., Holmes, W., Griffiths, M., & Forcier, L. B. (2016). Intelligence Unleashed: An Argument for AI in Education. Pearson.

Selwyn, N. (2020). Education and Technology: Key Issues and Debates. Bloomsbury Publishing.

Siemens, G., & Long, P. (2011). Penetrating the Fog: Analytics in Learning and Education. EDUCAUSE Review, 46(5), 30-32.
Wang, Y., Liu, C., & Zhang, X. (2021). Automating Education: The Role of AI in Distance Learning. Educational Technology & Society, 24(3), 12-25.

Zawacki-Richter, O., Marín, V. I., Bond, M., & Gouverneur, F. (2019). Systematic Review of Research on Artificial Intelligence Applications in Higher Education: Opportunities and Challenges. International Journal of Educational Technology in Higher Education, 16(1),

Laura Nemvalts
National Library of Estonia, Estonia

The Unfairness of OCR: The Example of Estonian Exile Newspapers

Thursday, 06/Mar/2025 2:20pm - 2:40pm
ID: 219 / Session SP 08: 2
Short paper (abstract) | 15-minute presentation with a 5-minute Q&A
Keywords: Estonian newspapers, Library materials--Digitization, Optical character recognition

The largest Estonian exile newspapers, published in Sweden, Canada, USA, Australia, and Germany, had been digitised by 2010 (Valmas, 2017). The earliness of this digitisation indicates the cultural significance of the material, as exile newspapers were important contributors to Estonian culture and politics during the Cold War period. However, over the past 15 years, digitisation, OCR, and segmentation tools have evolved rapidly. A contradictory situation has emerged: while these culturally significant newspapers were prioritised for early digitisation, their OCR quality is now significantly worse than that of more recently digitised materials. This leads to an unintended digital inequality, where researchers often prefer to use higher-quality digital sources, neglecting older yet culturally important material that has become less accessible due to poor OCR accuracy. The imbalance in the usability of digitised materials can cause research biases.

Memory institutions usually allocate resources to digitising materials that have not yet been digitised, rather than revisiting past digitisation projects. Thus, if digital humanities scholars still wish to study the material, which has been digitised some time ago, they inevitably face the question of whether they can rely on their research results. For researchers, the metadata of the digitisation process, including the date of digitisation and the average OCR accuracy, would be extremely helpful. As the documentation is often incomplete, especially in the case of earlier digitisation projects, it is difficult to assess or improve the reliability of the data.

OCR and segmentation errors are not new issues in the field of digital humanities, particularly in the study of historical newspapers. Oberbichler et al. (2021) conclude that materials digitised before the late 2010s often have insufficient OCR and segmentation quality, and the situation is further complicated by the lack of information about the digitisation process. According to Torget (2023), OCR accuracy is a key factor in assessing the usability of digitised newspapers. Several digital humanities projects have used historical newspapers as data, including NewsEye (newseye.eu, n.d.), which has also contributed to the re-OCRing of historical materials (Oberbichler et al., 2021).

In the process of improving the OCR quality, three types of OCR errors can be examined. First, the problems with OCR text can be caused by poor source material, for example, if the original newspaper text has faded, it becomes impossible to identify the text accurately. In these cases, re-digitising the higher-quality originals is often the only solution. Second, even when the source material is in good condition, illegible text can be caused by an outdated OCR model. This issue could be resolved by applying more reliable OCR tool to the same digital images. Third, even accurate OCR models can produce minor errors such as incorrectly identified characters. These can be corrected through OCR post-correction.

The presentation will explore potential approaches to compensate for missing metadata as well as how to improve the OCR that could be corrected with digital tools. In terms of OCR tools, this presentation will highlight the possibilities of the *Transkribus* platform in the case study focusing on Estonian exile newspapers published in Sweden in August 1991. In *Transkribus*, the NLF_Newseye_GT_FI_M2+ model was used, which was developed by the National Library of Finland during the Newseye project. This was further trained on Estonian-language material from the exile newspapers. With OCR post-correction, the potential of other tools, including artificial intelligence-based methods, will be explored. The preliminary results show that the use of such methods enhances the quality of newspaper metadata and OCR. However, for newspapers with complex layouts, improved OCR quality alone is not sufficient if there are many article segmentation errors. Thus, improved OCR does not guarantee improved text quality.

Digital humanities scholars have a vital role in improving the quality and usability of digitised sources. By utilising digital humanities tools and methods, they can help bridge the gap between older digitisation efforts and modern research needs. The impact of the research increases when its results are communicated back to GLAM institutions, where workflows can be adjusted to enhance the quality of digitised materials, which also benefits regular users by improving the functionality of digital archives.

Bibliography

newseye.eu. (n.d.). NewsEye: A Digital Investigator for Historical Newspapers. Retrieved January 31, 2025, from https://www.newseye.eu/

Oberbichler, S., Boroş, E., Doucet, A., Marjanen, J., Pfanzelter, E., Rautiainen, J., Toivonen, H., & Tolonen, M. (2021). 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. https://doi.org/10.1002/asi.24565

Torget, A. J. (2023). Mapping Texts: Examining the Effects of OCR Noise on Historical Newspaper Collections. In E. Bunout, M. Ehrmann & F. Clavert (Eds.), Digitised Newspapers – A New Eldorado for Historians?: Reflections on Tools, Methods and Epistemology (pp. 253–273). De Gruyter Oldenbourg. https://doi.org/10.1515/9783110729214

Valmas, A. (2017). Varjust valgusesse: erihoiust väliseesti kirjanduse keskuseks. Teaduste Akadeemia Kirjastus.

Gunta Nešpore-Bērzkalne, **Mikus Grasmanis**, **Lauma Pretkalniņa, Andrejs Spektors**
Institute of Mathematics and Computer Science, University of Latvia, Latvia

Enhanced digitized version of K. Mīlenbahs' “Dictionary of the Latvian Language”

ID: 175 / **Poster Session 2: 10**
Poster and demo (abstract) with accompanying a 1-minute lightning talk
Keywords: K. Mīlenbahs' “Dictionary of the Latvian Language”, digitization, electronic dictionary; legacy dictionary

"Dictionary of the Latvian Language" (Latviešu valodas vārdnīca) is a unique general scientific dictionary of the Latvian language, which was started in the 1880s by Kārlis Mīlenbahs (Mühlenbach), then edited, supplemented and completed by J.Endzelīns and E.Hauzenberga-Šturma (Kļaviņa 2023). It is both an explanatory and a translation dictionary with features of an etymology and synonym dictionary (Nešpore et al. 2006). In 1994, the digitization of the dictionary (Mīlenbachs 1923–1932) and its additional volumes (Endzelīns, Hauzenberga 1934–1946) was started at the IMCS, University of Latvia; in 2002, the electronic version of the dictionary (MEV) was published. It contains 132,718 entries.

In order to make the dictionary more accessible, in 2024, MEV was modernized and integrated with the Tēzaurs.lv platform (Grasmanis et al. 2023; https://mev.tezaurs.lv/).

The enhanced version provides several new and improved features:

- cross-dictionary search, both from MEV to other Tēzaurs dictionaries, and vice versa. On the MEV side, search works both on the original and the modern transcription of the entry words. In the opposite direction, the search function looks for entry words in modern transcription, and displays the original transcription;
- lexical neighborhood of the entry word;
- intra-dictionary cross-linking between entries, although limited by the pure presentational nature of the data markup;
- all entries for one entry word are shown together, ordered by homonym number, then main volumes before the additional volumes;
- links from entry to its facsimile page(s);
- improved appearance;
- previously missing materials (foreword, afterword, etc.);
- all diacritical marks have been converted to consistent Unicode symbols.

The markup files have been automatically checked for violations of markup syntax and structure, unrecognized markup tags, etc. Additionally, inconsistencies between the order of entries in the markup and the lexical order of entry words has been used as a criteria. All automated checks were followed by manual verification against the facsimile pages. Another source for verification was a manually created list of entry words, absent or undetectable (misspelled) in the electronic version.

Due to the dictionary's wide use of various languages, alphabets and additional diacritical marks (data of ~50 languages mentioned), MEV digitalization was an extreme challenge for OCR solutions of the early 2000s leading to a significant amount of OCR errors.

These automatic and manual verifications and corrections resulted in more than 23 thousand changes in the markup files.

During verification we detected entries with more than one entry word. In the old version, these additional entry words were ignored, or, in some cases, the whole entry was duplicated for each entry word. In the enhanced version, an additional lexeme has been added to the entry as a separate data element, duplicated entries were removed.

User feedback shows that the dictionary has become much more accessible. Cross-linking with the popular dictionary Tēzaurs has brought many new users. The large amount of fixed markup errors greatly improves the quality of the digitized dictionary.

Further steps include a transfer from presentational to semantic markup. However, this will be a challenge due to imprecise structures in the original and the remaining markup errors.

Bibliography

https://mev.tezaurs.lv/

Endzelīns, Jānis, Hauzenberga, Edīte. (1934–1946). Papildinājumi un labojumi K. Mülenbacha Latviešu valodas vārdnīcai. 1.–2. sēj. Rīga: Kultūras fonds.

Mīlenbahs, Kārlis. (1923–1932). Latviešu valodas vārdnīca. 1.–4. sēj. Rediģējis, papildinājis, turpinājis J. Endzelīns. Rīga: Izglītības ministrija, Kultūras fonds.

Grasmanis, Mikus, Paikens, Pēteris, Pretkalniņa, Lauma, Rituma, Laura, Strankale, Laine, Znotiņš, Artūrs, Grūzītis, Normunds. Tēzaurs.lv – the experience of building a multifunctional lexical resource. Electronic lexicography in the 21st century: Invisible Lexicography. Proceedings of the eLex 2023 conference, 2023, 400–418.

Kļaviņa, Sarma. “Latviešu valodas vārdnīca” (1923). Nacionālā enciklopēdija. Pieejams: https://enciklopedija.lv/skirklis/183820 "Latviešu-valodas-vārdnīca" (1923) [accessed 24.09.2024.].

Nešpore, Gunta, Grūzītis, Normunds, Andronova, Everita, Spektors, Andrejs. K. Mīlenbaha un J. Endzelīna Latviešu valodas vārdnīcas pilnveidota elektroniskā versija. Letonikas pirmais kongress. Valodniecības raksti. Rīga: Latvijas Zinātņu akadēmija, 2006, 241–249.

Kristoffer Nielbo
Center for Humanities Computing, Aarhus University, Denmark

Modeling Change: A Framework for Event Detection in Sociocultural Dynamical Systems

ID: 312 / WS04A: 5
Explorations of the dynamics of cultural phenomena in text corpora
Keywords: Change detection, Change description, Information theory, Digital history

A central building block for historical research is historical events, that is, dynamic objects displaced in time. Despite their importance, we see a disconnect between theoretical work and empirical studies of events [1]. This is exemplified by what we will refer to as the *Euclidean Error* in historical reconstructions. While historians generally agree that historical events are complex and non-linear in theory, empirical research is ripe with approaches that, due to data sparsity or inadequate formalization, describe history as consisting of singular dates, `event change points' that are connected by uneventful lines, static `event states' with low sensitivity to temporal variation, and, consequently, an overly reductive reconstruction of historical events. To counter this approach, we propose an alternative formal framework that offers an information-theoretical approach to historical event detection and description in noisy and complex sociocultural data. The framework is based on a fundamental theorem of chaos theory, the embedding theorem [2,3,4], which allows us to approximate the dynamics of a large-scale social system. Rather than measuring cultural expressions through, for instance, word counts over time, we approach culture as a complex system with a multitude of states, which switch between attractors, i.e., a value or set of values toward which variables in a dynamical system tend to evolve. Some of these attractors may be associated with the dynamics of cultural information and captured in low-dimensional indicator variables [5,6]. In our case studies, these indicator variables are expressed through the amount of surprise encoded in the textual content of news media. By this, we mean how much of the information at one point in time can be expected given an earlier time point. If the data is almost the same, there is a low level surprise; if it is radically different, the surprise increases. The framework is fundamentally data agnostic and will apply to any dense and low-rank embedding of the data objects, e.g., text, sound, or image, with some minor modifications. Importantly, our approach to events is psychological, i.e., we study how humans organize and understand events rather than attempt to formalize an event ontology [7]. The talk describes two techniques for detecting and describing changes between event states. However, the specific choice of models and algorithms is secondary to the main argument, namely that digital historical research has to pay more attention to complexities involved with change and events.

Bibliography

[1] Theo Jung and Anna Karla. 1. Times of the Event: An Introduction. History and Theory, 60(1):75–85, 2021.

[2] N. H. Packard, J. P. Crutchfield, J. D. Farmer, and R. S. Shaw. Geometry from a time series. Phys. Rev. Lett., 45(9):712–716, 1980. Publisher: American Physical Society.

[3] Floris Takens. Detecting strange attractors in turbulence. In David Rand and Lai-Sang Young, editors, Dynamical Systems and Turbulence, Warwick 1980, pages 366–381. Springer Berlin Heidelberg, 1981.

[4] Tim Sauer, James A. Yorke, and Martin Casdagli. Embedology. Journal of Statistical Physics, 65(3):579–616, 1991.

[5] Edward Ott. Chaos in Dynamical Systems. Cambridge University Press, 2 edition, 2002.

[6] Jianbo Gao, Yinhe Cao, Wen-wen Tung, and Jing Hu. Multiscale Analysis of Complex Time Series: Integration of Chaos and Random Fractal Theory, and Beyond. Wiley-Interscience, 1 edition edition, 2007.

[7] Antske Fokkens, Marieke Van Erp, Piek Vossen, Sara Tonelli, Willem Robert Van Hage, Luciano Serafini, Rachele Sprugnoli, and Jesper Hoeksema. GAF: A grounded annotation framework for events. In Workshop on Events: Definition, Detection, Coreference, and Representation, pages 11–20, 2013.

Ida Nordlander[1], **Alicia Fagerving**[2]
[1]Swedish Centre for Architecture and Design, Sweden; [2]Wikimedia Sverige, Sweden

From Data Cleanup to Linked Open Data: Hands-on with OpenRefine and Wikidata

ID: 199
Half-day conference-themed workshop
Keywords: Linked Open Data, OpenRefine, Wikidata, GLAM, Authority Control

A Network of Places: Open Linked (Building) Data as Research Infrastructure is a collaborative research and development project which aims to connect museum and archival collections using linked open data practices, initially looking at data related to the geographical areas Stockholm and Sápmi. The project is led by ArkDes (the Swedish Centre for Architecture and Design) and involves Nationalmuseum (the National Museum of Fine Arts), Tekniska museet (the National Museum of Science and Technology), the Swedish National Heritage Board's archive, and Wikimedia Sverige.

The challenge many cultural heritage institutions face today is the vast amount of data stored in their databases. Often this information remains inaccessible by the public and stored in closed systems. The lack of access to collection materials for researchers, the public and other cultural heritage institutions, also makes it difficult to analyze potential overlapping that occurs across collections. Therefore, the aim of the project is to demonstrate how museums and cultural heritage institutions alike can make their data more accessible to the public, and by doing so show the relevance of cultural heritage data by giving it further context.

ArkDes houses one of the world's largest architectural collections, featuring photographs, architectural drawings, models, and documents. This invaluable resource reflects the development of Swedish society through architectural material. In contrast, Nationalmuseum boasts a diverse array of collections that include not only architectural materials but also a rich variety of artworks, such as paintings, sculptures, and decorative arts. Tekniska museet has collections related to technology and development. Their collections focus on technological advancements and explore the theme of societal development, which ties in with the goals of this project. By enhancing existing data with geographical metadata, it would be possible to gain insights into the technological evolution of specific locations over time.

The Swedish National Heritage Board is Sweden’s central administrative agency in cultural heritage. Their archive houses records, drawings, photographs, and other documents related to archaeology, ancient relics, churches and other buildings, and cultural environments. The oldest documents in the archive date back to the 17th century. In addition to the physical archive, there is also a digital one based on the OAIS model. A very important part of the archive is the topographical series, in which the documents are organized according to their geographical location, which makes it possible to track what has happened at a specific location over time. Wikimedia Sverige, the Swedish chapter of the Wikimedia movement, has extensive experience supporting cultural heritage organizations in their work with free knowledge, including research projects of this kind being conducted through the Wikimedia platforms, including Wikidata. Wikidata has, prior to this project, been deployed as a central authority hub for cultural heritage data. Wikimedia Sverige provides a user’s perspective outside of the cultural heritage field. An example that illustrates this type of research is the three-year-long research and development project *Usable Authorities for Data-driven Cultural Heritage Research*. Whereas the current project focuses on building data and geographical metadata for linked open data practices, the prior project mainly focused on authoritative person data and utilized Wikidata as a platform for publishing and linking their cultural heritage data (Fagerving, 2023).

In this project, *A Network of Places: Open Linked (Building) Data as Research Infrastructure*, the foci lie in publishing geographical information as linked open data and demonstrating how such a method can be established as a commonly known practice within the cultural heritage sector – while establishing it as authority data. The core idea of applying linked open data methods, in this context, is to make cultural heritage more accessible, usable and easier to analyze. Imagine a researcher investigating a specific topic. Instead of having to consult various sources, all the information is collected in the same place – with additional linkages to related materials found in museum collections. It is reasonable to assume that there is some overlapping in the collected materials from each cultural heritage institution related to geographic information.

For example, while there is known overlap concerning the geographic area of Sweden, the vastness of the area makes detailed analysis challenging. Narrowing it down to a specific area or coordinate location would allow for more focused analysis of overlaps in materials, such as those related to a particular building and its geospatial data. Cultural heritage institutions today face the challenge of managing vast amounts of data stored in closed databases, making this information often inaccessible to the public. This lack of access hinders researchers, the public, and other institutions from analyzing potential overlaps across collections.

Having 100 objects from four institutions related to a specific city provides valuable insights. By linking 30 of these objects to the same location or building, we unlock new research opportunities. Utilizing linked open data allows us to show how a building connects to its municipality and country, ensuring the relevance of these objects. This approach helps us identify the 30 objects specifically associated with a particular building while also considering a broader collection across Sweden, thereby enhancing our understanding of their significance.

The workshop

The workshop will focus on the methods employed by the research group for identifying and enriching collection data with Wikidata. Participants will learn how to utilize the open-source tool OpenRefine to enhance their own collection data and actively contribute to improving existing data on Wikidata. This can be achieved by uploading external identifiers that link to their collections or by updating current metadata on the platform.

The target group for the workshop is mainly those who work with cultural heritage data. However, OpenRefine is a powerful tool that can be applied in a multitude of ways for data clean-up and analysis. The workshop can also be helpful for those who have previously worked with platforms such as Wikipedia and Wikidata and who want to hone their skills in the field. Participants do not have to have any prior experience with Wikidata, Wikipedia or OpenRefine, but a basic understanding of data management.

The workshop will be structured into two parts.

1. In the first part of the workshop participants will be informed about the on-going research project, useful terminology in linked open data practices and how to use Wikidata as a hub for linking and enriching data.
1. The second part of the workshop will be practical. Participants will get the opportunity to either bring their own data to clean, analyze and match to Wikidata through OpenRefine or use a test dataset provided by

the research team (this alternative is more suitable for beginners). The expected outcome of the practical part of the workshop is for participants to gain new insights into their own data and how linked open data can be applied to accomplish this. They will also gain an understanding of the Wikimedia open knowledge ecosystem.

The expected outcome from attending the workshop will be:

- A basic understanding on how to perform data analysis through OpenRefine.
- How to match collection metadata to Wikidata through OpenRefine.
- How to enrich your own data and metadata with Wikidata.
- How to upload and edit data to Wikidata through OpenRefine.
- An understanding of how to link data to Wikidata by using external identifiers.
- An understanding of the Wikimedia open knowledge ecosystem.

To-do before the workshop:

- **Bring a laptop.** You'll need a laptop to participate in the hands-on activities.
- **Register a Wikimedia user account.** If you don't have an account yet, please take a few minutes before the workshop to sign up. This will give you access to all the tools and resources we'll be using.
- **Download OpenRefine (optional).** To enhance your understanding of the tool, feel free to download OpenRefine ahead of time. It's a powerful data management tool that we'll be working with during the workshop. We'll also present an alternative way of using it on a cloud platform, which does not require you to install anything locally.
- **Bring your own data (optional).** If you have specific data you'd like to work on, please bring it along! This can make the exercises more relevant and tailored to your needs.

Bibliography

Fagerving, A. (2023). Wikidata for authority control: sharing museum knowledge with the world. In Digital Humanities in the Nordic and Baltic Countries Publications (Vol. 5, Issue 1, pp. 222–239). University of Oslo Library. https://doi.org/10.5617/dhnbpub.10665

Fredrik Nylén, Tomas Skotare, Johan von Boer
Humlab, Umeå University, Sweden

The Visible Speech (VISP) platform: A secure infrastructure for the study of speech acts and spoken conversations

ID: 202 / Poster Session 2: 19
Poster and demo (abstract) with accompanying a 1-minute lightning talk
Keywords: digital speech signal processing, secure research platform, national collaboration support

Visible Speech (VISP) is a web-based research infrastructure at Humlab, Umeå University, designed to handle audio recordings of speech in compliance with the national implementation of GDPR and security requirements. VISP provides a centralised environment for research of all disciplines in which recordings of spoken language constitute the primary material, meeting both researchers' needs for efficient workflows and legislators' demands for secure data management.

One of VISP's primary advantages is its ability to facilitate research on audio recordings that constitute personally identifiable information (PII) under Swedish law. These recordings may further contain sensitive content or have been made in sensitive contexts, classifying them as sensitive PII under national legislation. Sensitive contents may occur in relation to, for instance, the ethnicity and religious beliefs of the speaker, and sensitive contexts may occur when the recording is made in a healthcare context or in a context where a person's membership with a union organisation is divulged.

The VISP platform offers a unified environment for storage, controlled access, direct work, and reproducible speech signal processing. It includes the most comprehensive set of speech and voice analysis procedures available within one framework globally. Additionally, VISP facilitates the digital archiving of projects through a uniform, documented, and transparent directory structure, reducing barriers to making data available in accordance with the FAIR principles. Research projects dealing with sensitive personal data in audio recording form require review by the Ethical Approval Authority and may subsequently take advantage of the VISP facilities.

A significant feature of VISP is its integration with the Swedish Academic Identity Federation (SWAMID) which enables secure, federated login for researchers across Sweden. This national federated login system allows researchers to access project data and collaborate on material processing in ways that were previously not possible. Moreover, VISP supports projects by lowering the step in to digital signal processing and audio analysis of the collected audio signals. This capability allows researchers to perform hands-on processing and analysis without the risk of disseminating sensitive audio recordings. By leveraging SWAMID, VISP ensures that researchers can work seamlessly and securely on collected materials, enhancing collaborative efforts and data handling efficiency. By providing tools for direct manipulation and examination of audio data, VISP ensures that all stages of data handling, from collection to analysis, are conducted within a secure environment, thereby maintaining the integrity and confidentiality of sensitive information.

The work conducted within VISP is part of SweCLARIN, the Swedish node of the European Research Infrastructure Consortium (ERIC) CLARIN. SweCLARIN aims to develop and provide national and European infrastructure for speech and text-based e-science, offering extensive digitized materials and advanced language technology tools. By combining advanced technology with stringent security protocols and leveraging national federated login systems, VISP enables efficient and secure research on audio recordings of speech. This makes it an invaluable tool for researchers, facilitating unprecedented collaboration and data processing within the digital humanities.

Niklas Nylund
Vapriikki Museum Centre, Finland

Safekeeping the demoscene together with the community

ID: 293 / **WS02: 3**
Innovations and New Interactions through Digital Cultural Heritage in GLAM Sector
Keywords: demoscene, digital heritage, intangible heritage, collaboration, safekeeping, preservation

This paper presents the results of collaborative projects in the Finnish Museum of Games and Finnish Postal Museum to safekeep the Finnish demoscene. The projects have focused on 'non-traditional' GLAM methods, such as workshops and film documentaries, in order to 1) enable the continuity of the demoscene culture, 2) make born-digital heritage, often perceived as 'difficult' due to its 'exclusive' and 'technical' nature, more accessible to new audiences, and 3) amplify new and alternative perspectives in the demoscene community.

The demoscene is a grassroots sub-culture of creative computing that sprung into existence in the late 80s and early 90s all around Europe. Its origins are in the networks and values surrounding software piracy and cracker counterculture of the mid-80s (Albert 2017; Reunanen 2014). Demos are real-time computer programs, often including graphics and music, that push the limits of the hardware they run on. Making demos form the core of the demoscene culture, but it is the diverse social networks and their shared values and skills that have become of interest to academic researchers and GLAM institutions (e.g. Nylund & Suvanto 2023; Albert 2020).

Finland was the first country in the world where the demoscene was incorporated on the National List of Intangible Heritage in 2020. This admission has made it possible to fund preservation projects in Tampere in collaboration with demoscene members. In the 2023 project Demoskenen uudet kasvot ("New Faces of the Demoscene"), workshops dedicated to the intergenerational transfer of skills related to the demoscene were organised. The Demoskene talteen ("Preserving the Demoscene") project in 2024 in turn produced a 30- minute video documentary tailing the experiences of three teenagers coming into contact with the demoscene.

In these projects, the demoscene has been re-interpreted. As the demoscene has traditionally seen meritocratic competition as its core value, it has focused only on preserving the digital artefacts it produces, instead of focusing on the people, skills, values and experiences behind demoscene culture. The workshops and video documentary produced have highlighted how the community has begun conscious efforts to make the demoscene more accessible. The projects have given a voice to otherwise silenced minorities inside the community and the original meritocratic values of the demoscene have been reinforced by collaborative and emphatic approaches, which are helping to safekeep the intergenerational continuity of demoscene. These visions welcome collaboration with museums in building up the demoscene as sustainable form of born-digital culture.

Bibliography

Albert, Gleb J. 2017. From Currency in the Warez Economy to Self-Sufficient Art Form: Text Mode Graphics and the 'Scene'. WiderScreen 20 (1-2). http://widerscreen.fi/numerot/2017-1-2/from-currency-in-the-warez-economy-to-selfsufficient-art-form-text-mode-graphics-and-the-scene/.

Albert, Gleb J. 2020. NewScenes, New Markets: The Global Expansion of theCracking Scene, Late 1980sto Early 1990s. WiderScreen 23 (2-3). http://widerscreen.fi/numerot/2020-2-3/newscenes-new-markets-the-globalexpansion-of-the-cracking-scene-late-1980s-to-early-1990s/.

Nylund, Niklas & Suvanto, Eljas (2023). Demoskene (digitaalisena) kulttuuriperintönä ("The Demoscene as (Digital) Cultural Heritage"). Suomen Museo-Finskt Museum, 130, 27-45.

Reunanen, Markku. 2014. How Those Crackers Became Us Demosceners. WiderScreen 17 (1-2). http://widerscreen.fi/numerot/2014-1-2/crackers-became-usdemosceners/

Dalia Ortiz Pablo, Maria Skeppstedt, Anna Foka
Centre for Digital Humanities and Social Sciences, Dept. of ALM, Uppsala University, Sweden

A Cross-University Collaborative Approach for Python Course Development: Observations from a Digital Humanities Perspective

ID: 291 / WS09: 2
Workshop on Digital Humanities and Social Sciences/Cultural Heritage (DHSS/DHCH) in Higher Education
Keywords: Course development, Python programming, Inter-university cooperation

The Wallenberg AI and Transformative Technologies Education Development Program (WASP-ED) addresses the critical need for scalable, timely education in AI and transformative technologies across Sweden. In particular, WASP-ED's Work Area 3: Course Development focuses on cross-sectoral and inter-university cooperation to develop AI education, with a particular emphasis on Python programming. Through a structured learning pathway, the initiative seeks to equip students and professionals with the necessary skills to harness changing technologies effectively in their fields. This effort aligns with broader European and global initiatives to address the increasing need for upskilling and reskilling in transformative technologies.

In the course development package, the Centre for Digital Humanities and Social Sciences at Uppsala University (CDHU), together with Lund, Umeå and Luleå Universities, has developed a series of foundational Python courses. The series starts at the beginners' level, continues with a deeper exploration of standard and external library packages, and ends with a focus on code quality and object-oriented programming. In this workshop presentation, we discuss the development of the second-level course - Programming in Python: Standard and External Library Packages. In this regard, our presentation has three main objectives. First, we aim to showcase the practical administrative complexities in work carried out to ensure a smooth transition for students across universities and courses, highlighting the close collaboration in course content development and administrative coordination. Second, we elaborate on the development of hands-on teaching materials, including pre-recorded lectures, recommended readings, laboratory exercises, quizzes, and practical programming assignments. Finally, we demonstrate how CDHU leverages its expertise in digital humanities and social sciences to create relevant course content. In conclusion, we exemplify how this initiative may be used as a blueprint to highlight the power of cross-sectoral collaboration in technological education, bringing together digital humanities and programming to address the growing demand for technical competence across disciplines. By combining the strengths of multiple universities and incorporating real-world applications from digital humanities, we show how the program offers a unique and comprehensive learning experience.

Michael Thomas Leonce Pace-Sigge
University of Eastern Finland, Finland

Large-Language-Model Tools and the Theory of Lexical Priming: convergence and divergence of concepts of language

Thursday, 06/Mar/2025 10:00am - 10:30am
ID: 165 / Session LP 06: 3
Long paper (full-text) | 20-minute presentation with a 10-minute Q&A

Keywords: (Artificial Intelligence (AI), Human Cognition, Large Language Models (LLMs), Lexical Priming Theory, Naturally Occurring language, Spreading Activation

This paper revisits Michael Hoey’s *Lexical Priming Theory* (2005) in the light of recent discussions of *Large Language Models* forms of machine learning (commonly referred to as AI) which have found a lot of publicity in the wake of tools like OpenAI’s *ChatGPT* or Google’s *BARD/Gemini*. Historically, theories of language faced inherent difficulties, given language's exclusive use by humans and the complexities involved in studying language acquisition and processing. The intersection between Hoey's theory and Machine Learning tools, particularly those employing Large Language Models (LLMs), has been highlighted by several researchers. Hoey's theory relies on the psychological concept of priming, aligning with approaches dating back to Ross M. Quillian's 1960s proposal for a "Teachable Language Comprehender." The theory posits that every word is primed for discourse based on cumulative effects, a concept mirrored in how LLMs are trained on vast corpora of text data.

This paper tests LLM-produced samples against naturally (human-)produced material in the light of a number of language usage situations, investigates results from A.I. research and compares the results with how Hoey describes his theory. While LLMs can display a high degree of structural integrity and coherence, they still appear to fall short of meeting human-language criteria which include grounding and the objective to meet a communicative need.

Michael Thomas Leonce Pace-Sigge
University of Eastern Finland, Finland

What a Nano-GPT can (not) tell us about Spoken Language

ID: 170 / Poster Session 2: 28
Long paper (full-text) | 20-minute presentation with a 10-minute Q&A
Keywords: Generative Pre-trained Transformer (GPT), Large Language Model, nano-GPT, Scouse, spoken English, word clusters

After the launch of Chat-GPT in autumn 2022, a lot of research has focussed on the quality and near-naturalness which Large Language Model (LLM)-based tools present in the texts produced. While one area of research focussed on the similarities and differences between machine-produced and human-produced output (e.g. Berber Sardinha, 2024), others explored in how far such tools could process more complex tasks (e.g. Valmeekam, et al., 2023; Curry et al., 2024.

While it can be assumed that Chat-GPT makes use of written-to-be-spoken training material, there has been no investigation, as yet that looks at in how far a Generative Pre-trained Transformer (GPT) algorithm is able to process (transcribed) natural, colloquial language. This research will investigate whether spoken language transcripts lead to processing difficulties; whether such generated language can be seen as a suitable reflection of natural speech; and whether machine produced texts offer new insights into the workings of language.

Olha Petrovych[1,2], Liina Saarlo[1], Mari Väina[1], Kaarel Veskis[1,3]
[1]Estonian Literary Museum, Estonia; [2]Vinnytsia Mykhailo Kotsiubynskyi State Pedagogical University, Ukraine; [3]University of Tartu, Estonia

From Oral Tradition to Digital Insights: A Comparative Topic Analysis of Ukrainian and Estonian Folk Songs

Thursday, 06/Mar/2025 3:40pm - 4:00pm
ID: 184 / Session SP 07: 6
Short paper (abstract) | 15-minute presentation with a 5-minute Q&A
Keywords: Thematic Analysis, Topic Modeling, Latent Dirichlet Allocation, Estonian and Ukrainian Folk Songs, Comparative Analysis, Digital Humanities, Translation.

Folk songs, as carriers of cultural memory, encapsulate historical events, social norms, moral values, and emotional experiences within a community. Building on Jan Assmann's (2011) theory of cultural memory, which distinguishes between "communicative memory" (short-term, based on personal recollections) and "cultural memory" (long-term, institutionalized through texts, rituals, and artifacts), this study explores how computational analysis can reveal the structures and themes that constitute the cultural memory encoded in folk songs.

This research applies computational methods to analyse the thematic structures of Ukrainian and Estonian folk songs, utilizing latent Dirichlet allocation (LDA) topic modelling – a computational technique that identifies latent topics based on statistical word co-occurrence within texts. The aim is to examine how themes, motifs, and narrative structures are distributed across these two distinct folk traditions and to assess the feasibility of LDA for analysing both corpora. Additionally, the study seeks to enhance traditional folkloristic analysis by incorporating computational methods, including translation with large language models (LLMs) and topic modeling techniques (Blei, Ng, & Jordan 2003; Egger 2022; Griffiths & Steyvers 2004; Sarv 2020), to provide a comprehensive comparison of thematic structures.

Oral tradition theory (Parry 1971; Lord 1991) emphasizes the formulaic nature of oral composition and the role of tradition in shaping narrative content. This theory provides a foundation for understanding how repeated phrases, motifs, or thematic structures aid in recall and convey cultural meanings (e.g., Foley 1988; 1991; 1995; Frog & Lamb 2022). Building on this theoretical background, the current study uses topic modeling to uncover recurring themes and motifs within a large corpora of Ukrainian and Estonian folk songs, revealing how these elements function across different cultural contexts.

In addition, this research also aligns with the FILTER project (2020–2024), which has developed innovative computational tools for the analysis of complex linguistic materials, which are applicable not only to Finnic oral poetry but also to folk songs of other nations.

The research addresses three primary questions: (1) What underlying themes, motifs, and narrative structures can be identified within Ukrainian and Estonian folk songs using topic modelling? (2) How do the thematic structures derived from computational analysis align with traditional folkloristic classifications? (3) How does the use of translation impact the analysis of thematic overlap between these two languages?

The research is based on two main corpora: Ukrainian folk songs from the Podillia region collections (Dei 1965; Iefremova & Dmytrenko 2014; Myshanych 1976); and a corpus of Estonian folk songs (ERAB) maintained by the Estonian Literary Museum (Sarv & Oras 2020), with a focus on the songs from Järvamaa, as the dialects of this region are among the closest to the Estonian written language.

The text data are subjected to LDA topic modeling to identify and analyse thematic clusters and narrative structures. The findings reveal distinct thematic clusters within each language's folk songs, highlighting both shared motifs and unique cultural elements.

To ensure consistent comparison, the study explores the possibility of translating both corpora into English using LLMs and analysing the translated texts for thematic overlap. This approach aims to provide deeper insights into how themes in Ukrainian and Estonian folk songs align or diverge, and to determine the impact of translation on thematic analysis.

The comparative analysis of thematic structures between Ukrainian and Estonian folk songs is conducted based on the following criteria:

- Thematic overlap: Identifying common themes across both languages and examining the degree of thematic overlap.

- Cultural context: Analysing how cultural and linguistic differences influence thematic expression and motif representation in the songs.

- Topical divergence: Evaluating unique themes or motifs specific to each folk tradition.

- Semantic clusters: Comparing semantic clusters derived from topic modeling to traditional folkloristic classifications to assess alignment and divergence.

- Statistical comparison: Utilizing statistical measures to quantify the similarities and differences in thematic distribution between the two corpora.

By combining LDA topic modeling with the analysis of translated texts using LLMs, the study uncovers distinct thematic clusters within each language's folk songs, highlighting both shared motifs and culturally specific elements. By applying topic modeling to Ukrainian and Estonian folk songs, the study offers a quantitative framework that enriches folkloristic research, advancing the integration of digital methods in folkloristics and offering new perspectives on folk song traditions.

Bibliography

Assmann, Jan. 2011. Communicative and Cultural Memory. In: Meusburger, P., Heffernan, M., Wunder, E. (eds) Cultural Memories. Knowledge and Space, vol 4, pp. 15–27. Springer, Dordrecht. https://doi.org/10.1007/978-90-481-8945-8_2

Blei, David M., Ng, Andrew Y., & Jordan, Michael I. 2003. Latent Dirichlet Allocation. Journal of Machine Learning Research, 3, 993–1022.

Dei, Oleksii (red.). 1965. Pisni Yavdokhy Zuikhy: zapysav Hnat Tantsiura [Songs of Yavdokha Zuikha: recorded by Hnat Tantsiura]. Kyiv : Naukova dumka. 810 s.

Egger, Roman. 2022. Topic modelling. Modelling hidden semantic structures in textual data. In: Egger, R. (eds) Applied Data Science in Tourism. Tourism on the Verge. Springer, Cham. https://doi.org/10.1007/978-3-030-88389-8_18

FILTER – Formulaic intertextuality, thematic networks and poetic variation across regional cultures of Finnic oral poetry. Available at https://blogs.helsinki.fi/filter-project/, last accessed on 21 October 2024.

Foley, John Miles. 1988. The Theory of Oral Composition: History and Methodology. Bloomington.

Foley, John Miles. 1991. Immanent Art: From Structure to Meaning in Traditional Oral Epic. Bloomington.

Foley, John Miles. 1995. The Singer of Tales in Performance. Bloomington.

Frog & William Lamb (eds.). 2022. Weathered Words. Formulaic Language And Verbal Art. Cambridge, Massachusetts & London, England: Harvard University Press.

Griffiths, Thomas L., & Steyvers, Mark. 2004. Finding scientific topics. Proceedings of the National Academy of Sciences, 101(Suppl 1), 5228–5235. https://doi.org/10.1073/pnas.0307752101

Lord, Albert Bates. 1991. Epic Singers and Oral Tradition. Ithaca.

Myshanych, Stepan (red.). 1976. Pisni Podillia: zapysy Nasti Prysiazhniuk v seli Pohrebyshche. 1920-1970 rr. [Songs of Podillia: recordings of Nastia Prysiazhniuk in the village of Pohrebyshche. 1920-1970.] Kyiv: Naukova dumka. 520 p.

Parry, Milman. 1971. Making of Homeric Verse: The Collected Papers of Milman Parry. Ed. by Adam Parry. Oxford.

Sarv, Mari 2020. Regilaulude teema-analüüs: võimalusi ja väljakutseid [Topic analysis of Estonian runosongs: Prospects and challenges]. Methis. Studia Humaniora Estonica 21(26): 137–160. https://doi.org/10.7592/methis.v21i26.16914

Sarv, Mari; Oras, Janika. 2020. From tradition to data: The case of Estonian runosong. In: Arv. Nordic Yearbook of Folklore, 76, 105−117.

Yefremova, Liudmyla, & Mykola Dmytrenko (red.). 2014. Narodna pisni Khmelnychchyny (z kolektsii zbyrachiv folkloru) [Folk songs of Khmelnytskyi region (from the collections of folklore collectors)]. Kyiv: Naukova dumka. 720 s.

Chantal Pivetta[1], **Renato Caenaro**[3], **Ilaria Papa**[2]
[1]Lund University (Sweden), Sweden; [2]Università di Padova, Italy; [3]SilentWave SRLS

Show me your data! Visualization and interpretation of data from Cultural Heritage

ID: 274 / **Panel 1: 1**
60-minute panel
Keywords: data visualization, data analysis, linked data, the semantic web, and relational databases

Introduction

Researching, preserving, visualizing, and analyzing heritage assets is inherently complex, requiring a multidisciplinary approach. A core challenge in Cultural Heritage (CH) is not only safeguarding physical structures but also effectively communicating the significance and values they embody. Heritage is more than landscapes, buildings, or museum sites; it comprises both tangible and intangible values shaped by historical processes that define collective identity. This aligns with principles like those in the Faro Convention, emphasizing the connection between heritage, communities, and democratic values, often fostering active public involvement in conservation and promotion efforts.

Effectively conveying the multifaceted nature of CH demands rigorous research and interdisciplinary collaboration across fields such as IT, architecture, art history, archaeology, and engineering. By adopting structured methodologies and advanced tools, professionals ensure CH is studied holistically, preserving it as a living testament to cultural identity while engaging contemporary societies.

Linked Data and Integrated Systems: technological advancements have transformed CH research by enhancing data acquisition and enabling seamless integration of multidisciplinary insights. Linked data and interoperable systems connect material, documentary, and contemporary datasets, offering a unified perspective on the evolution of CH over time. These systems foster a relational approach, addressing the fragmented nature of traditional research.

Digital Tools and Collaboration: digital tools like Geographic Information Systems (GIS), web-based platforms, and 3D modeling technologies are reshaping CH workflows. Linked data systems streamline data collection, enable global collaboration, and support real-time information sharing. These methodologies enhance researchers' ability to handle complex datasets, opening new possibilities for architectural and archaeological studies.

Computational Methods to Bridge Gaps: one key challenge in CH research is the limited use of computational methods for analyzing historical records. Traditional methods often lack efficiency and integration, particularly for documentary sources. Tools such as Building Information Modelling (BIM), Historic Building Information Modelling (HBIM), and TEI-encoded datasets address these gaps by enabling systematic analysis of CH data. For example, studying the Benedictine Cassinese Congregation's architectural network (15th–18th centuries) highlights how reforms initiated in 1418 influenced the circulation of ideas, models, and practices within the Cassinese network. Using linked systems, this research enhances understanding of these monasteries as interconnected nodes rather than isolated sites.

Data Visualization in CH Research: data visualization plays a vital role in representing the spatial, temporal, and relational dimensions of CH. Advanced visualization techniques uncover patterns and relationships in datasets, facilitating deeper insights. Influential works, such as Edward Tufte's *The Visual Display of Quantitative Information*, and Digital Humanities scholars like Matthew K. Gold and Lauren F. Klein, underscore the importance of visualization in bridging computational methods and humanistic inquiry. Intuitive visual tools allow for the exploration of archives, maps, and museum collections, enriching CH research and engagement.

3D Modeling and Immersive Technologies: 3D modeling and virtual reconstructions are crucial for studying CH, providing immersive experiences that deepen understanding of historical spaces and urban design. Projects such as *Rome Reborn*, which digitally reconstructs ancient Rome, demonstrate how 3D technologies bridge past and present. Similarly, Jeffrey Schnapp's work at the Stanford Humanities Lab highlights the importance of reconstructing spatial and temporal dimensions in CH research. By combining linked data with immersive tools, these technologies make heritage more accessible and engaging for both scholars and the public.

Challenges and Emerging Directions

This panel aims to present three contributions from three different backgrounds, where stakeholders and researchers joined their forces address their specific goals. These studies illustrate how digital tools can transform the analysis of CH, enabling more thorough examinations of documentary and material sources. These examples highlight the power of digital strategies, methodologies and technologies to integrate diverse data sets, offering a more nuanced, comprehensive and multi-layered view of historical cultural heritage.

However, these case studies also face challenges related to data visualisation and analysis, particularly concerning the interpretation of information gathered from architectural and archaeological sources. As digital tools become more sophisticated, the challenge of presenting complex data in a way that is both accurate and accessible becomes increasingly important. The integration of new visualisation methods must balance technical precision with the need for clarity, ensuring that insights gained from digital models are effectively communicated to a wide range of audiences.

Data-Driven Heritage: Visualizing and Analysing the Architectural Heritage of the Benedictine Cassinese Congregation by Ilaria Papa, University of Turin

This paper focuses on the integration of Digital Humanities (DH) with computational methods to explore monastic architecture and historical documentation. This study is part of Ilaria Papa's PhD research, which examines the architectural heritage of the Benedictine Cassinese Congregation between the 15th and 18th centuries. It aligns with the PRIN CoenoBIuM project, applying BIM (Building Information Modelling) and HBIM (Historic Building Information Modelling) for digital and spatial analysis, and specifically seek to develop new strategies for the digital transition in historical-architectural research and for studying monastic architecture.

The research addresses a significant gap in the use of computational methods for analysing documentary sources related to architectural history. While BIM and HBIM have revolutionised how architectural heritage is digitally represented, the interpretation of historical documents remains largely dependent on traditional methods. This reliance often leads to inefficiencies and fragmentation in comparative studies, which require the ability to analyse complex interconnections between multiple sites.

By focusing on the Cassinese Congregation, the study examines how the Benedictine Order's reform initiated by Ludovico Barbo in 1418 led to the constitution of a real network of monasteries and produced tangible effects on the material aspects of the architecturesDespite this interconnectedness, traditional studies have often focused on individual sites rather than the broader cultural network. The project aims to fill this gap by examining the circulation of ideas, models, and practices within the Cassinese cultural network, utilizing TEI-encoded transcriptions of General Chapter reports as digital datasets. This is further supported by Edition Visualization Technology (EVT) and a relational database model for analysing metadata.

Textual Data Structured According to TEI Guidelines for Analysis and Visualization in Multidisciplinary Projects by Chantal Pivetta, University of Lund.

The digitization and structured analysis of historical-documentary sources are vital to advancing Cultural Heritage (CH) research. These processes provide unprecedented access to textual data, supporting complex relational analyses of interconnected corpora. This contribution presents a workflow and two specialized tools for managing textual data in CH projects, focusing on transcription, encoding, visualization, and analysis using Text Encoding Initiative (TEI) guidelines.

The workflow consists of two interconnected phases, with this presentation focusing on the first:

1. **Semantic Annotation and Encoding:** Documentary sources are collaboratively transcribed and encoded using TEI guidelines within a shared, user-friendly environment. Encoded data is stored in a relational database, enabling systematic organization and query-based retrieval.
2. **Complex Data Analysis and Integration:** Structured data supports advanced relational and geospatial analyses, linking textual data to external platforms like mapping tools.

Tools Supporting the Workflow:

Digital Philology for Dummies (DPhD): This user-friendly software simplifies TEI encoding, especially for participants with minimal XML/TEI experience. It provides:

- A shared workspace for consistent workflows and automated generation of shared lists and @xml:id values.
- Flexibility for transcribers with varying expertise, allowing tailored contributions.

Textual data encoded through this process can be analyzed using the integrated relational database and visualized with EVT.

Edition Visualization Technology (EVT): An open-source platform for visualizing XML/TEI-encoded texts. EVT's synoptic interface enables comparative and relational analysis, revealing insights into CH networks such as monastery studies. It facilitates navigation of encoded data and links textual information to external datasets.

Key Contributions

This case study highlights the value of TEI encoding and integrated tools in CH research by:

- Promoting data interoperability for seamless reuse.
- Supporting collaborative workflows for multidisciplinary teams.
- Enabling advanced relational and geospatial analysis.

By enhancing accessibility and engagement with historical data, the project sets a benchmark for interdisciplinary CH methodologies and advances the principles of open science.

Why Establishing a Data-Sharing Environment in Multidisciplinary Research Projects by Renato Caenaro, CEO of SilentWave SRLS.

The relational database serves as the technical backbone and core of the project mentioned in the first contribution, enabling seamless interoperability between diverse datasets, programming languages, and digital tools. It is designed to handle the complex requirements of cross-disciplinary research by supporting structured and semi-structured data workflows.

A semi-structured data-sharing environment is integral to this setup, providing researchers with a framework to follow clear guidelines and utilize tools for collecting and organizing data from multiple sources right from the start.

This environment ensures:

- Data consistency through uniform entry protocols,
- Alignment of team efforts by standardizing input workflows, and
- Simplified aggregation of shared insights, regardless of the contributors' disciplinary backgrounds.

Key Technical Features:

1. Automated Relationship Generation
   The environment supports the semi-automatic generation of relationships between datasets. This is achieved through predefined rules and dynamic linking facilitated by the database schema, enabling researchers to conduct cross-domain queries efficiently.
2. Integration with Open Data Standards
   The database adheres to openData and openAccess principles by enabling seamless publication of data using structured openAPIs. These APIs ensure that:

- Data can be exported as complete database dumps for full-scale analysis.
- Data can also be accessed as interconnected and linked datasets, enhancing interoperability with external systems.

Scalable Querying and Relationship Management
Advanced query optimization within the relational database supports complex interactions between datasets, allowing for efficient navigation of multidimensional relationships. The system facilitates bidirectional linking between textual, geospatial, and architectural data, creating a robust platform for advanced analytical workflows.

Benefits in Practice

This technical setup enables:

- Simplified data collection: Uniform rules streamline the input process, reducing errors and improving efficiency.

- Collaborative data sharing: A shared, multimodal environment allows researchers from different disciplines to contribute seamlessly.
- Automated linkage: Relationships between diverse datasets are generated programmatically, drastically reducing the manual workload.
- Cross-domain querying: Researchers can perform advanced searches and analyses across textual, geographical, and architectural data.
- Early-stage data publication: With APIs and open data principles integrated into the workflow, datasets are prepared for immediate publication and external use.

By incorporating these features, the relational database ensures a scalable, transparent, and interoperable solution for managing the complexities of this multidisciplinary research project (and potentially others). This approach not only simplifies collaboration across diverse teams but also aligns with the highest technical standards for open data and linked data interoperability.

Bibliography

Andrews, T. L. (2013). The Third Way: Philology and Critical Edition in the Digital Age. The Journal of the European Society for Textual Scholarship, 10, 61–76. https://doi.org/10.1163/9789401209021_006

Baker C., Cheung K.-H. (eds) (2007). Semantic Web—Revolutionizing Knowledge Discovery in the Life Sciences. Berlin/Heidelberg: Springer.

Bleier, R., Bürgermeister, M., Klug, H. W., Neuber, F., & Schneider, G. (Eds.). (2018). Digital Scholarly Editions as Interfaces (Vol. 12). BoD. https://kups.ub.uni-koeln.de/9085/

Burghart, M. (2017). Creating a Digital Scholarly Edition with the Text Encoding Initiative. https://halshs.archives-ouvertes.fr/halshs-01984319

Cheng G., Shao F., Qu Y. (2017). An empirical evaluation of techniques for ranking semantic associations. IEEE Transactions on Knowledge and Data Engineering, 29(11): 1.

Ciotti, F., Corradini, E., Cugliana, E., D'Agostino, G., Ferroni, L., Fischer, F., Lana, M., Monella, P., Roeder, T., Turco, R. R. D., & Sahle, P. (2022). Manifesto per le edizioni scientifiche digitali. Umanistica Digitale, 12, Article 12. https://doi.org/10.6092/issn.2532-8816/14814

Council of Europe. (2005). Faro Convention. https://www.coe.int/it/web/venice/faro-convention

Council of Europe. 2000. European Landscape Convention. http://www.coe.int/en/web/landscape/home.

de Laat, R., and L. Van Berlo. (2011). "Integration of BIM and GIS: The Development of the CityGML GeoBIM Extension." In Advances in 3D Geo-Information Sciences, 211–225.

Di Pietro, C., & Rosselli Del Turco, R. (2018). Between Innovation and Conservation: The Narrow Path of User Interface Design for Digital Scholarly Editions. In Digital Scholarly Editions as Interfaces (Vol. 12, pp. 133–163). BoD. https://kups.ub.uni-koeln.de/9115/

Eero, H. Heikki, R. (2021). Knowledge-based relational search in cultural heritage linked data, Digital Scholarship in the Humanities, Volume 36, Issue Supplement_2, Pages ii155–ii164, https://doi.org/10.1093/llc/fqab042

Favretto, A. (2022). Relational Database, GIS Layers, and Geodatabase for Cultural Heritage Management In: D'Amico, S., Venuti, V. (eds) Handbook of Cultural Heritage Analysis. Springer, Cham. https://doi.org/10.1007/978-3-030-60016-7_46.
Fischer, M., and M.B.S. Salih. (2020). "Data-Driven Cultural Heritage: A Data Science Perspective." Heritage Science, 8(1), 1–15. https://doi.org/10.1186/s43238-020-00002-1

Giordano, A., and P. Borin. (2016). "HBIM: Analisi critica tra didattica e ricerca." In Brainstorming BIM. Il modello tra rilievo e costruzione, 53–63. Milan.

Hyvönen, E. (2020). Using the Semantic Web in Digital Humanities: shift from data publishing to data-analysis and serendipitous knowledge discovery. Semantic Web, 11(1):187–93.

Hyvönen E. (2012). Publishing and Using Cultural Heritage Linked Data on the Semantic Web. Palo Alto, CA: Morgan & Claypool.

Indrawan-Santiago, M. (2012). Database research: Are we at a crossroad? Reflection on NoSQL. In 2012 15th International Conference on Network-Based Information Systems (pp. 45-51). IEEE.

Khalil, A. M., and K. Arslan. (2022). "The Role of Digital Humanities in Cultural Heritage Research." Journal of Digital Humanities, 6(3), 245–260.

McCarty, W. (1999). Humanities Computing as Interdiscipline. http://www.iath.virginia.edu/hcs/mccarty.html [Accessed: 19 Oct. 2024]

Monella, P., & Rosselli Del Turco, R. (2020). Extending the DSE: LOD Support and TEI/IIIF Integration in EVT. In C. Marras, M. Passarotti, G. Franzini, & E. Litta (Eds.), Atti del IX Convegno Annuale dell'Associazione per l'Informatica Umanistica e la Cultura Digitale (AIUCD). La svolta inevitabile: Sfide e prospettive per l'Informatica Umanistica (pp. 148–155). Available online as a supplement of Umanistica Digitale: https://umanisticadigitale.unibo.it. https://doi.org/10.6092/UNIBO/AMSACTA/6316

Münster, S., M. Pfarr-Harfst, P. Kuroczyński, and M. Ioannides, (eds. 2016). 3D Research Challenges in Cultural Heritage II: How to Manage Data and Knowledge Related to Interpretative Digital 3D Reconstructions of Cultural Heritage. Lecture Notes in Computer Science, vol. 10025. Cham, Switzerland: Springer.Robinson, P. M. W. (2022). An approach to complex texts in multiple documents. Digital Scholarship in the Humanities, fqab108. https://doi.org/10.1093/llc/fqab108

Rosselli Del Turco, R. (2019). Designing an advanced software tool for Digital Scholarly Editions: The inception and development of EVT (Edition Visualization Technology). Textual Cultures, 12(2), 91–111. https://doi.org/10.14434/textual.v12i2.27690

Roberto Rosselli Del Turco, Giancarlo Buomprisco, Di Pietro Chiara, Kenny Julia, Masotti Raffaele, & Pugliese Jacopo. (2015). Edition Visualization Technology: A Simple Tool to Visualize TEI-based Digital Editions. Journal of the Text Encoding Initiative, Issue 8, 1–21. https://doi.org/10.4000/jtei.1077

Schreibman, S., & Papadopoulos, C. (2019). Textuality in 3D: Three-dimensional (re)constructions as digital scholarly editions. International Journal of Digital Humanities, 1(2), 221–233. https://doi.org/10.1007/s42803-019-00024-6

Tamper M., Leskinen P., Apajalahti K., Hyvönen E. (2018). Using biographical texts as linked data for prosopographical research and applications. In Digital Heritage. 7th International Conference on Progress in Cultural Heritage: Documentation, Preservation, and Protection, EuroMed 2018, Springer, Nicosia, Cyprus.

Tommasi, C., C. Achille, and F. Fassi. (2016). "From Point Cloud to BIM: A Modelling Challenge in the Cultural Heritage Field." International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences, 41, 429–436.

Vincent, M., F. Kuester, and T. Levy. (2013). "OpenDig: In-field Data Recording For Archaeology And Cultural Heritage." In Proceedings of the IEEE Conference Digital Heritage, Marseille, France, 28 Oct–1 Nov, Vol. 1, edited by A.C. Addison, L. De Luca, G. Guidi, and S. Pescarin, 539–542. Piscataway, NJ: Institute of Electrical and Electronics Engineers.

von Schwerin, J., M. Lyons, L. Loos, N. Billen, M. Auer, and A. Zipf. (2016). "Show Me the Data!: Structuring Archaeological Data to Deliver Interactive, Transparent 3D Reconstructions in a 3D WebGIS." In 3D Research Challenges in Cultural Heritage, edited by S. Münster, M. Pfarr-Harfst, P. Kuroczyński, and M. Ioannides. Lecture Notes in Computer Science, vol. 10025. Cham, Switzerland: Springer.

Chantal Pivetta[1], Roberto Rosselli Del Turco[2], Renato Caenaro[3]
[1]Lund University (Sweden); [2]Università di Torino, Italy; [3]Silentwave SRLS

Create and Showcase Your Digital Critical Edition: A Step-by-Step Guide with Digital Philology for Dummies (DPhD) and Edition Visualization Technology (EVT)

ID: 186
Full-day conference-themed workshop
Keywords: XML TEI coding, authoring editor, digital scholarly editions, editions visualization technology, coding and transcribing texts

Introduction of the workshop topic/idea and its importance:

Digital Scholarly Editions (DSEs) are essential for the preservation and accessibility of cultural heritage, providing global access to rare and historically significant texts. They support advanced research by offering searchable, annotated, and – in some cases – interlinked texts that clearly represent textual variations and editorial decisions. By integrating digital humanities tools, DSEs enable new methods of analysis and encourage interdisciplinary collaboration. Their interactive features and – when available – high-resolution images of original manuscripts make complex materials more engaging and useful for both research and teaching. As a result, DSEs ensure that historical and literary works remain accessible, transparent, and relevant in the digital era.

This workshop highlights the importance of encoding manuscripts and visualizing them using specialized digital tools. This process is crucial for preserving, analyzing, and making complex textual data more accessible. Encoding captures detailed information about a manuscript's variations, structure, and materiality in a machine-readable format, improving searchability and enabling advanced analysis. Visualization tools then transform this encoded data into an interactive platform, allowing users to explore patterns, conduct comparative analyses, and understand editorial decisions. This approach enhances both scholarly research and public engagement, offering new ways to study and appreciate cultural and historical texts.

The workshop guides participants through the complete process of creating a Digital Scholarly Edition, focusing on two main phases with integrated software tools that simplify the workflow. In the first phase, participants will transcribe and encode manuscripts or printed texts using DPhD (Digital Philology for Dummies). The second phase involves visualizing the completed edition with Edition Visualization Technology (EVT). This hands-on approach allows participants to work directly with digital tools and materials, resulting in a practical project they can use as a foundation for their own research. By the end of the workshop, participants will have developed a digital edition that can be further refined and expanded according to their research needs.

The software that will be used:

Digital Philology for Dummies (DPhD) is a user-friendly software designed to facilitate the transcription and encoding of texts, regardless of their nature or language. It speeds up the process for experienced users while allowing beginners to focus on their humanistic research objectives without needing to worry—at least initially—about more technical or technological aspects.

Edition Visualization Technology (EVT) plays a crucial role in the visualization of Digital Scholarly Editions (DSEs), offering specialized tools for presenting and analyzing complex textual data. EVT enhances accessibility through a user-friendly interface, enabling users to navigate layered information and compare textual variants effectively. By integrating high-quality images with textual transcriptions, EVT allows for a richer understanding of manuscripts, preserving the original context and materiality of texts. It also promotes transparency in editorial practices, providing insights into the decision-making process behind transcriptions and annotations. The interactivity of EVT supports deeper engagement in both academic research and educational settings, while fostering collaboration and community participation.

Target audience:

The target audience includes anyone interested in digital scholarly editions or in the transcription of texts in XML/TEI format as well as their visualization and navigability. Beginners eager to learn are welcome, as are experts who can benefit from the more advanced and new features of the tools presented.

Expected outcomes:

Participants in this workshop will gain practical skills in transcribing, encoding, and visualizing digital texts using Digital Philology for Dummies (DPhD) and Edition Visualization Technology (EVT). They will complete a prototype digital edition that can serve as a foundation for future projects and acquire a holistic understanding of the Digital sScholarly Edition (DSE) workflow.

In detail:

- Learn how to transcribe and encode texts, including manuscripts or printed materials, using Digital Philology for Dummies (DPhD). This skill is essential for creating accurate digital editions that retain the textual and structural nuances of original documents.
- Familiarity with TEI Standards: Gain a foundational understanding of the Text Encoding Initiative (TEI) standards, which are widely used in digital humanities for encoding textual features. This knowledge will be valuable for future digital editing projects.
- Mastery of Edition Visualization Technology (EVT): Become proficient in using EVT to display and interact with encoded texts. Develop skills such as integrating transcriptions with high-quality images, visualizing textual variants, and navigating through different layers of the text.
- Creating User-Friendly Digital Editions: By the end of the workshop, you will know how to produce digital editions that are both accurate and accessible, making them ideal for academic research and teaching.
- A Practical Digital Edition: Work hands-on with provided materials and software throughout the workshop, resulting in a practical digital edition that you can take away as a prototype. This outcome can serve as a basis for creating more advanced editions or incorporating your own research materials.
- Application to Future Projects: Reflect on how the methods and tools used in the workshop can be applied to your research projects, helping you plan and implement your own digital scholarly editions.

Bibliography

Apollon, Daniel, and Claire Belisle. 2014. Digital Critical Editions. University of Illinois Press.

Buzzoni, M., & Rosselli Del Turco, R. 2024. “Towards an Integrated Digital Edition of the Leges Langobardorum”. In Atti del XIII convegno AIUCD “Me.Te. Digitali: Mediterraneo in rete tra testi e contesti”. Retrieved from https://iris.unive.it/handle/10278/5078701.

Driscoll, Matthew James, and Elena Pierazzo, eds. 2016. Digital Scholarly Editing: Theories and Practices. Open Book Publishers. http://www.openbookpublishers.com/product/483.

Edition Visualization Technology. 2014-. Home page: https://evt.labcd.unipi.it. GitHub repository: https://github.com/evt-project/evt-viewer-angular/.

Rosselli Del Turco, Roberto. 2017. The Digital Vercelli Book. A Facsimile Edition of Vercelli, Biblioteca Capitolare, CXVII. Collane@unito.It. Università di Torino. https://www.collane.unito.it/oa/items/show/11.

Rosselli Del Turco, Roberto. 2019. “Designing an Advanced Software Tool for Digital Scholarly Editions: The

Marina Platonova, Tatjana Smirnova, Zane Seņko, Oksana Ivanova
Riga Technical University, Institute of Digital Humanities, Latvia

Preserving Authenticity in Postgraduate Education: Authorship Identification vs Artificial Intelligence

Wednesday, 05/Mar/2025 4:20pm - 4:40pm
ID: 182 / Session SP 04: 2
Short paper (abstract) | 15-minute presentation with a 5-minute Q&A
Keywords: Authenticity, artificial intelligence, authorship identification, text-generation technology, transmedia storytelling

A prominent essay by Roland Barthes claiming the death of the author stirred considerable turmoil back in 1967, opening the ground for a heated discussion of such issues as authorial intention, the role of the reader as a (co)creator of the text, and the concept of authorship *per se*. The latter being of crucial importance in the educational space since within the academic environment the concepts of unblemished authorship identification, academic integrity, ethical research, and plagiarism are being put into question when the personal contribution of an individual to the creation of authentic content may not be accurately measured and appreciated. Although information mining and clustering have long become automated and specialists with various backgrounds have appreciated the operational simplicity achieved, nowadays digital technologies are being increasingly used rather in information synthesis and presentation if not content creation. Emerging trends and practices of AI-aided content generation have repeatedly raised concerns with regard to the recognition, attribution, and ethical use of human-generated intellectual property and artifacts. It is hard to underestimate how generative AI tools being widely used in the academic setting, have changed the paradigm of learning and have especially influenced postgraduate performance assessment metrics, as the master level of higher education addresses a relevant number of research aims. At the postgraduate level, students demonstrate an ability to critically assess the existing theoretical framework to be able to synthesize new knowledge, present and substantiate their opinions and findings. Hence, they are expected to produce a considerable volume of written and oral output that is both relevant and authentic.

Universities worldwide have yet to decide on the policies to deal with the omnipresent, often intentionally malicious use of text-generation software in completing a range of tasks – essays, various study papers, and even master theses. The majority of effective codes of academic integrity or codes of ethics employed by higher education institutions imply immediate imposition of various sanctions for the use of content-generation software without exhibiting a proper acknowledgment of the fact, and often failing to timely react at the changing philosophy of authorship. The mechanism of sanctioning for unauthorized use of AI text-generation tools should be balanced with a range of educational activities aimed at promoting academic integrity and Master student awareness of the advantages and limitations these tools offer. Text-generation, text-processing, and digital storytelling tools may become both helpful and perilous in postgraduate research endeavors since their use may either improve or undermine academic performance.

The paper considers written output production habits and ethical attitudes of postgraduate students of the program “Digital Humanities” as well as the choice they have to make, addressing both the dichotomy of exposing creativity and preserving authorship versus producing task-rooted content with the help of various AI-technologies, and the juxtaposition of the overreliance on the technology with the value of authenticity and ingenuity. Ethical and sustainable use of generative AI tools is analyzed and discussed with the students, considering such issues as ethics and aesthetics of borrowing, reconsidering and adopting ideas expressed by others, critical thinking, and awareness of the notion of individual creativity, authenticity, and genuineness of own artistic, academic or scientific output expanding the borders of human knowledge.

Special focus is made on how principles, tools, and software allowing students to engage in transmedia storytelling are used to promote a conscious and ethical use of modern text-generation technology. In view of the fact that transmedia storytelling is inherently a creative process, it may encourage students and young researchers to explore various perspectives on their material. Students explore the scope of transmedia storytelling, which expands through the “Snowballing Effect”, and learn how culture enriches through reconsideration and reinterpretation of well-known stories.

Hence, the aim of the research is to explore the scenarios of authorized, sustainable, and ethical use of text-generation software in the university setting, consider the concept of authorship and academic integrity in view of the changing paradigm of postgraduate performance assessment, and discuss solutions and practices adopted in this area by the faculty of the study program “Digital Humanities” implemented by RTU.

Bibliography

Dudacek, O. (2015). Transmedia Storytelling in Education. Procedia – Social and Behavioral Sciences, 197, 694–696.

Freeman, M. (2016). Historicising Transmedia Storytelling: Early Twentieth-Century Transmedia Story Worlds. Routledge.

Hovious, A., Shinas, V. H., & Harper, I. (2021). The Compelling Nature of Transmedia Storytelling: Empowering Twenty First-Century Readers and Writers through Multimodality. Technol. Knowl. Learn. 26, 215–229. doi: 10.1007/s10758-020-09437-7

Jenkins, H. (2006). Convergence Culture: Where Old and New Media Collide. NYU Press.

Ryan, M.-L. (2013). Transmedial Storytelling and Transfictionality. Poetics Today, 34(3), 361–388.

Alois Andreas Põdra[1], Anni Martin[2], Üllar Alev[3]
[1]Tallinn University of Technology // Estonian Open Air Museum; [2]Ministry of Climate, Estonia; [3]Tallinn University of Technology // Heritage Board of Estonia

Digital renovation passport: a new approach to the preservation of heritage homes

Thursday, 06/Mar/2025 3:20pm - 3:40pm
ID: 239 / Session SP 08: 5
Short paper (full-text) | 15-minute presentation with a 5-minute Q&A
Keywords: Building renovation passport, renovation wave, energy efficiency, heritage buildings, digitizing heritage

The heritageHOME project focuses on digitizing the renovation process and solutions for heritage homes. These buildings usually have very poor energy performance in their original condition, and the road to an energy-efficient home is much more challenging.

Built heritage is a cornerstone of the European identity and has an important role in creating and preserving community values and a sense of belonging. But heritage buildings are also homes – the people who live there also want modern comforts, well-maintained properties, and affordable housing. Neglecting these buildings in favor of widespread renovation can contribute to the risk of abandonment and vacancy. To tackle this issue, Estonia has taken on the ambition of attempting to consciously address energy-efficient solutions for heritage buildings.

HeritageHOME project is ongoing, and the first step of digital tool development has been to map out a universal, yet comprehensive renovation roadmap of heritage homes, including common problems, necessary stages, likely risks, involved constituencies, etc. The method involved both workshops with specialists and interviews with homeowners. While the first iteration of the tool can be described as a digital knowledge repository, the goal is to compile an interactive platform utilizing the previously mentioned universal renovation roadmap phases and renovation solutions to generate tailored renovation roadmaps with interdisciplinary cooperation.

Estonians, as digital natives, are therefore taking digitalization of built heritage to the next level – by expanding and applying it to the service of heritage buildings!

Henna Poikkimäki[1], Petri Leskinen[1,2], Eero Hyvönen[1,2]
[1]Aalto University, Finland; [2]University of Helsinki, Finland

Exploring Cultural Heritage Knowledge Graphs—Case of Correspondence Networks in Grand Duchy of Finland 1809–1917

Wednesday, 05/Mar/2025 2:00pm - 2:30pm
ID: 211 / Session LP 01: 2
Long paper (full-text) | 20-minute presentation with a 10-minute Q&A
Keywords: digital humanities, network analysis, correspondence networks, linked data, knowledge graphs

This paper argues for using methods and tools of Network Analysis (NA) to study contents of knowledge graphs (KG) in Digital Humanities (DH) research. As a case study, social and correspondence networks in the Grand Duchy of Finland 1809–1917 are considered with a focus on prosopographical data about historical people and, in particular, their correspondences (epistolary data).

Letters have been an important form of communication, and networks based on letter metadata, letter's content, and related biographical information can be used for rebuilding and analyzing historical social networks and for studying the flow of ideas and information. In correspondence network analysis, ego-networks focusing on only one person and his correspondents are common due to the nature of letter collections. Combining letter collections and biographical data helps move from ego-centric network approach towards sociocentric networks, as the larger network starts to emerge when letter collections from many individuals are brought together, although analyses still suffer from missing data. In this paper, we present results of the Constellations of Correspondence (CoCo) project that so far has created a KG of over million letters exchanged during 1809–1917 in Finland, reusing data of prosopographical KGs of the same period of time.

Ajda Pretnar Žagar[1], Rajko Muršič[2]
[1]Faculty of Computer and Information Science, University of Ljubljana, Slovenia; [2]Faculty of Arts, University of Ljubljana, Slovenia

Sensory-digital explorations of urban ambiances

ID: 286 / **WS09: 5**
Workshop on Digital Humanities and Social Sciences/Cultural Heritage (DHSS/DHCH) in Higher Education
Keywords: interdisciplinary teaching, computational anthropology, sensory walk, urban ambiances, sensory anthropology

The contribution presents an interdisciplinary approach to teaching anthropology, showcasing how computational methods can enhance anthropological inquiry (Pretnar Žagar and Podjed 2024). The project-based learning methodology introduced students to sensory walks (Abram 2023, Järviluoma 2022), inviting them to capture images of their surroundings while engaging their sensory perceptions of urban environments and ambiances. The goal was to explore what guided their sensory interest during the walk (Muršič 2019).

The student-collected images were analyzed using data science techniques (Paff 2022), specifically hierarchical clustering with image embeddings, cosine distance, and Ward linkage. The analysis identified two distinct clusters: one featuring nature and the other urban environments. Surprisingly, both clusters were consistently present across all participants, indicating that students were equally drawn to urban structures and natural elements, regardless of their environment.

This study underscores the potential for integrating qualitative anthropological methods with computational tools to uncover patterns in human perception (see also Beaulieu 2017). The interdisciplinary approach not only deepened students' understanding of anthropology but also demonstrated the value of blending qualitative fieldwork with computational analysis (Bornakke and Due 2018) as well as experimenting with digital devices as the extension of human sensoria (Podjed and Muršič 2021). It provided students with hands-on experience in combining ethnographic sensitivity with machine learning, fostering a holistic view of their surroundings.

At the workshop, we will further illustrate this approach by inviting participants to undertake a blended sensory walk in Tartu, practically testing the relation between sensorial and digital (Muršič 2021). Their images will be analyzed using the described methodology, allowing attendees to experience firsthand how anthropology and data science can intersect to provide rich insights into cultural and environmental perceptions. This model offers a novel framework for teaching and research, highlighting the relevance of computational techniques in addressing anthropological questions.

Bibliography

Abram, S., 2023. Sensoryfication of place: A sensobiographic approach to sensing transformations of urban atmospheres. In: Ambiance, Tourism and the City. Routledge, pp. 137–148.

Beaulieu, A., 2017. Vectors for fieldwork: Computational thinking and new modes of ethnography. In: Hjorth, L., Horst, H., Galloway, A., and Bell, G., eds. The Routledge Companion to Digital Ethnography. New York; London: Routledge, pp. 55–65.
Bornakke, T. and Due, B.L., 2018. Big – Thick Blending: A method for mixing analytical insights from big and thick data sources. Big Data & Society, 5(1), pp. 1–16. https://doi.org/10.1177/2053951718765026.

Järviluoma, H., 2022. Sensobiographic walking and ethnographic approach of the Finnish school of soundscape studies. In: The Bloomsbury handbook of popular music, space and place, pp. 83–98.

Jenks, A., Lowman, C., and Straughn, I., 2024. AI for Learning: Experiments from Three Anthropology Classrooms. Anthropology News. Available at: https://www.anthropology-news.org [Accessed 10 December 2024].

Muršič, R., 2019. Sensory walking: teaching methods in motion. Teaching Anthropology: A Journal of the Royal Anthropological Institute. Available at: https://www.teachinganthropology.org/2019/01/31/sensory-walking-teaching-methods-in-motion/ [Accessed 10 December 2024].

Muršič, R., 2021. Between aisthēsis and colere: sensoria, everyday improvisation and ethnographic reality. Amfiteater: Revija za teorijo scenskih umetnosti, 9(2), pp. 134–153. Available at: https://www.slogi.si/wp-content/uploads/2021/12/Amfiteater_9_2_Raz_09_Mursic_EN.pdf [Accessed 10 December 2024].

Paff, S., 2022. Anthropology by Data Science. Annals of Anthropological Practice, 46, pp. 7–18. https://doi.org/10.1111/napa.12169.

Podjed, D. and Muršič, R., 2021. To be or not to be there: remote ethnography during the crisis and beyond. Etnolog, 82(31), pp. 35–51. Available at: https://www.etno-muzej.si/files/etnolog/pdf/0354-0316_31_Podjed_Mursic_To.pdf [Accessed 10 December 2024].

Pretnar Žagar, A. and Podjed, D., 2024. Ethnography beyond thick data. Annals of Anthropological Practice, 48, pp. 272–288. https://doi.org/10.1111/napa.12226.

Pille Pruulmann-Vengerfeldt[1], Pille Runnel[1], Kai Pata[2], Mahendra Mahey[3]
[1]Estonian National Museum, Estonia; [2]Tallinn University, Estonia; [3]Strathclyde University, Scotland UK

Digital cultural heritage as a resource for social development

ID: 214 / **Panel 3: 1**
60-minute panel
Keywords: digital cultural heritage, GLAM labs, use cases

A significant part of Estonian museum collections is available digitally. It is an invaluable social resource with great potential to contribute to sustainable development. The rapid social, economic and technological development leads to polarization and a sense of discontinuity, but heritage can help people cope with change. Museums help to cope with transition when their resources find application in society. Therefore, the panel focuses on the application of digital cultural heritage. The panel discussions showcase how digital heritage can make sense and contribute to social development. The presentations discuss the theoretical and analytical frameworks, categorisation and classification of digital heritage use and practical examples of how heritage has been put to use. The examples discussed in the panel presentation touch local and international use cases.

Paper 1: Information environment model to understand pain points in putting digital cultural heritage in service for social development

How does individual use of digital cultural heritage bring about social change? Presenting a preliminary analytical framework, the presentation looks at different steps from availability, relevance, and use to cultural change, trying to illustrate how creative projects that put heritage to use can result in social change. The information environment framework originally presented by Leah Lievrouw (2001) allows us to showcase that the uses of digital heritage are bound both by the institutional as well as personal/relational aspects. This means that digital heritage can be used if there is institutional availability, accessibility, awareness of tradition and practice of putting digital heritage to use, but also to the social context of the user, their understanding of relevance, skills they have or that they can lean on, and willingness to put in effort. While there are cases where the individual heritage use might be outstanding, the question is if the effort expected from the individual is proportionate to the perceived use and value of it. In many cases, digital heritage becomes useful and usable in the context of the heritage institution, where professionals become mediators for the digital cultural heritage. However, the current digital heritage is often far from accessible (even if it is digitally available) for the regular user, as even motivated specialists struggle to access knowledge digitised in the databases. The knowledge dissemination framework allows us to identify particular pain points that currently hinder using digital cultural heritage for social development and provides an analytical framework to discuss where the least amount of effort might get the most significant results.

Paper 2: Analytic categories for mapping digital heritage use cases: building a framework for understanding knowledge transformation and impact

In the research project „Digital cultural heritage as a social resource“ the research team explores how to systematically assess and compare diverse digital heritage use cases to better understand their properties and role in knowledge transformation processes for broader societal impact.

As the use of digital heritage continues to expand, encompassing a wide range of cases from creative and artistic uses of digital archives to virtual exhibitions, there is a growing need for structured approaches to analyze, evaluate, compare these varied applications. This presentation introduces how the research team is developing a systematic framework that incorporates a set of analytical categories to assess the properties of digital heritage use cases and applications and their role in the knowledge transformation process. The framework is designed to help evaluate use cases by focusing on their innovation potential, user engagement, technical and organizational requirements, and potentially their contribution to societal impact, as framed by Pier Luigi Sacco’s Culture 3.0 impact areas.

The framework builds on a set of analytical categories developed through qualitative data collection and mapping from various digital heritage projects, covering contextual, data, and technology-related factors, as well as the agency of users, including user interaction and skills.

This presentation will discuss how these categories can be operationalized into a tool that enables the comparative analysis of digital heritage use cases, supporting more informed decision-making in the GLAM sector. This spans from design and content-related decisions to strategic development, offering a holistic way to evaluate digital heritage applications.

By examining use cases through this lens, we aim to support future digital heritage projects, emphasizing collaboration, participation, and social innovation. Ultimately, this framework has the potential to help practitioners understand how their initiatives can transition digital heritage from static repositories to dynamic tools for social change.

Paper 3: How can digital heritage support social and cultural resilience?

This presentation discusses the critical points in the digital cultural heritage ecosystem from the point of view of building cultural resilience states - coping, adapting and transforming. The analytical view of digital heritage use considering some Estonian GLAM ecosystem examples highlights these critical points in which the cultural ecosystem dynamics are harmed. The awareness of the digital heritage in the cultural ecosystem for specific communities, and the scaffolding aspects in formal and informal learning processes directed at building knowledge, values, identity and place belonging are discussed from the perspective of cultural resilience and adaptations.

Paper 4: Understanding how the diverse uses of digital cultural heritage in British Library Labs and the other Gallery, Library, Archive and Museum (GLAM) Labs could teach us about designing activities and services to facilitate social development

Provisional findings of interviews conducted with users of digital cultural heritage largely at the British Library will be highlighted, including patterns identified when mapping responses onto an information environment model developed by Lievrouw (2001) and expanded by Pille Pruulmann-Vengerfeldt (2024). The analysis will focus on various factors including what was done with digital heritage and how, what skills were required to implement the projects, how access was provided, who benefited from the work, what (if any) impact was achieved and whether it resulted in social development etc.

There will be an analysis of interviews with those who have been involved in providing services to enable users to work with digital cultural heritage in GLAM experimental physical and/or virtual spaces (Labs). There will be a special focus on how these services evolved and changed over time, especially in the context of changing institutional priorities (for example after COVID) and the increasing interest in Artificial Intelligence.

There will be a discussion as to the provisional findings of this work and what components could be included in future work to enable social development when working with digital cultural heritage.

Bibliography

Liewrouw, L. A. (2001). New Media and the `Pluralization of Life-Worlds': A Role for Information in Social Differentiation. New Media & Society, 3(1), 7-28. https://doi.org/10.1177/1461444801003001002

Anna Puhakka
Finnish National Gallery, Finland

Art and Technology in Harmony: A Case Study of the Finnish National Gallery's Combine24 Competition

ID: 299 / **WS02: 5**
Innovations and New Interactions through Digital Cultural Heritage in GLAM Sector
Keywords: generative art, digital cultural heritage, innovations, co-creation, blockchain

In 2024, the Finnish National Gallery issued a challenge to the blockchain-based art community: engage with its copyright-free CC0 collection through Combine24, an international generative art competition. This initiative exemplifies how technology and art can unite to reinterpret cultural heritage and foster innovative creative practices.

This presentation will explore the Combine24 competition as a case study, offering valuable insights for cultural institutions and creatives alike. It will detail the competition’s structure, objectives, and outcomes, shedding light on how museums can serve as active facilitators of artistic innovation. By inviting artists to interact with collection databases and associated metadata, the Finnish National Gallery opened its archives to new possibilities, positioning itself at the forefront of digital creativity and cultural engagement.

Key themes of the presentation include:

1. **Museums as Collaborative Partners:** How institutions can bridge traditional and contemporary practices by encouraging interaction between historical collections and cutting-edge technology.
1. **Technology as a Creative Catalyst:** The potential of generative art and blockchain platforms to reinterpret and expand the boundaries of cultural heritage.
1. **Fostering Interdisciplinary Connections:** Insights into how collaborations between technologists, artists, and cultural

organizations can lead to meaningful and unexpected outcomes.

The session will illustrate how Combine24 successfully leveraged the FNG CC0 collection and metadata, inspiring artists to create works that blend historical and contemporary narratives. By encouraging creative engagement, the competition served as a model for how museums can transform static collections into dynamic resources for dialogue and innovation.

Through this case study, attendees will gain practical strategies for fostering similar collaborations, utilizing open-access collections, and integrating technology to build new pathways for interaction with cultural heritage. Ultimately, the presentation underscores the transformative potential of digital tools in reimagining the role of museums in the 21st century.

This session invites cultural institutions, technologists, and creatives to envision a future where art and technology work hand-in-hand to preserve the past, while simultaneously creating new forms of interaction and storytelling for global audiences.

Heikki Rantala[1], **Eero Hyvönen**[1,2], **Eljas Oksanen**[3,2], **Jouni Tuominen**[2]
[1]Aalto University, Finland; [2]University of Helsinki, Finland; [3]University of Reading, UK

Opening Archaeological Public Finds Data with Semantic Web Technologies: Demonstrating FindSampo, CoinSampo, and PASampo

ID: 212 / **Poster Session 2: 23**
Poster and demo (abstract) with accompanying a 1-minute lightning talk
Keywords: linked data, semantic web, archaeology, citizen science

This demonstration presents three Linked Open Data (LOD) services and semantic portals for searching, exploring, and analyzing data related to collections of archaeological public finds in Finland and UK.

1. FindSampo - Finnish Archaeological Public Finds on the Semantic Web} (in use at https://loytosampo.fi) is based on the finds collection of the Finnish Heritage Agency (Rantala et al. 2022).
2. CoinSampo - Finnish Numismatic Public Finds 2013-2023} (in use at https://rahasampo.ldf.fi/) (Rantala et al. 2024) is based on the systematic recordings of numismatic finds in Finland at the National Museum of Finland.
3. PASampo is a related, as of yet unpublished, system based on the database of over one million finds of the Portable Antiquity Scheme in England and Wales at the British Museum (https://https://finds.org.uk/) (Lewis et al. 2024).

It is shown, using practical examples, how the underlying SPARQL endpoint and the semantic portals built on top of the data services available at the Linked Data Finland platform (https://ldf.fi) (Hyvönen at al. 2014; Hyvönen and Tuominen 2024) can be used in opening Cultural Heritage (CH) data. Our aim is also to demonstrate the "FindSampo framework" (Hyvönen et al. 2021) for publishing archaeological public finds developed based on the Sampo model (Hyvönen 2022). The examples also demonstrate the use of Sampo-UI framework (Ikkala et al. 2022; Rantala et al. 2023) in developing new applications for Digital Humanities research.

The web applications offer especially users with limited technical background, such as metal detectorists, an easy way to visualize the data with various charts and maps. The demonstrators also show how LOD can be used to enrich data, for example by fetching coordinates for places, such as minting places of coins, from outside sources to be visualized on maps. With the demonstrator applications, users can through combination of facet selections and visualizations, for example, browse finds from a certain munipality, visualize minting places or find spots of coins from a certain period on a map, or compare relative numbers of objects made from different materials using bar charts.

The LOD services and portals demonstrated are part of the Sampo series of over 20 systems (https://seco.cs.aalto.fi/applications/sampo/) published across 20 years 2004-2024 for opening CH data.

Bibliography

Hyvönen, Eero. 2022. "Digital Humanities on the SemanticWeb: Sampo Model and Portal Series." Semantic Web 14 (4): 729–744. https://doi.org/10.3233/SW-223034.

Hyvönen, Eero, Heikki Rantala, Esko Ikkala, Mikko Koho, Jouni Tuominen, Babatunde Anafi, Suzie Thomas, et al. 2021. "Citizen Science Archaeological Finds on the Semantic Web: The FindSampo Framework." Antiquity, A Review of World Archaeology 95, no. 382 (August): e24. https://doi.org/10.15184/aqy.2021.87.

Hyvönen, Eero, and Jouni Tuominen. 2024. "8-star Linked Open Data Model: Extending the 5-star Model for Better Reuse, Quality, and Trust of Data." In Posters, Demos, Workshops, and Tutorials of the 20th International Conference on Semantic Systems (SEMANTiCS 2024), vol. 3759. CEUR Workshop Proceedings. https://ceur-ws.org/Vol-3759/paper4.pdf.

Hyvönen, Eero, Jouni Tuominen, Miika Alonen, and Eetu Mäkelä. 2014. "Linked Data Finland: A 7-star Model and Platform for Publishing and Re-using Linked Datasets." In The Semantic Web: ESWC 2014 Satellite Events, 226–230. Springer-Verlag, May. https://doi.org/10.1007/978-3-319-11955-7_24.

Ikkala, Esko, Eero Hyvönen, Heikki Rantala, and Mikko Koho. 2022. "Sampo-UI: A Full Stack JavaScript Framework for Developing Semantic Portal User Interfaces." Semantic Web 13 (1): 69–84. https://doi.org/10.3233/SW-210428.2

Lewis, Michael, Eljas Oksanen, Frida Ehrnsten, Heikki Rantala, Jouni Tuominen, and Eero Hyvönen. 2024. "The Impact of Human Decision-making on the Research Value of Archaeological Data." Submitted for evaluation (June). https://seco.cs.aalto.fi/publications/2024/lewis-etal-pasampo-2024.pdf.

Rantala, Heikki, Annastiina Ahola, Esko Ikkala, and Eero Hyvönen. 2023. "How to create easily a data analytic semantic portal on top of a SPARQL endpoint: introducing the configurable Sampo-UI framework." In VOILA! 2023 Visualization and Interaction for Ontologies, Linked Data and Knowledge Graphs 2023. CEUR Workshop Proceedings, Vol. 3508. https://ceurws.org/Vol-3508/paper3.pdf.

Rantala, Heikki, Esko Ikkala, Ville Rohiola, Mikko Koho, Jouni Tuominen, Eljas Oksanen, Anna Wessman, and Eero Hyvönen. 2022. "FindSampo: A Linked Data Based Portal and Data Service for Analyzing and Disseminating Archaeological Object Finds." In The Semantic Web: ESWC 2022, 13261:478–494. Lecture Notes in Computer Science. Springer. https ://doi.org/10.1007/978-3-031-06981-9_28.

Rantala, Heikki, Eljas Oksanen, Frida Ehrnsten, and Eero Hyvönen. 2024. "Publishing Numismatic Public Finds on the Semantic Web for Digital Humanities Research – CoinSampo Linked Open Data Service and Semantic Portal." In SemDH 2024, First International Workshop of Semantic Digital Humanities, co-located with ESWC 2024, Hersonissos, Greece, Proceedings. CEUR Workshop Proceedings, May. https://ceur-ws.org/Vol-3724/paper3.pdf.

Maciej Rapacz, Aleksander Smywiński-Pohl
AGH University of Kraków, Poland

Ancient Text, Modern Methods: Transformers in Interlinear Translation

Wednesday, 05/Mar/2025 3:00pm - 3:30pm
ID: 197 / Session LP 02: 4
Long paper (abstract) | 20-minute presentation with a 10-minute Q&A
Keywords: natural language processing, machine translation, low resource settings, ancient greek

# Introduction

Machine Translation (MT) has seen significant advancements in recent years, particularly with the introduction of transformer architectures (Vaswani et al., 2017). While MT typically aims for natural and fluent translations, there exists a spectrum of translation approaches, ranging from free translation to extremely literal renderings. At the far end of this spectrum lies interlinear translation (Shuttleworth and Cowie, 2014), an approach that prioritizes preserving the original syntactic structure over fluency. This method, frequently (although not exclusively) used on sacred texts, arranges target language words directly below or above their corresponding source text items. Interlinear translation serves as a linguistic bridge, providing access to texts for those who may lack the necessary language skills to approach them directly. As Gutt (1991) points out, this approach aims to preserve not only lexical equivalence but also syntactic categories such as word order with minimal changes. Consequently, interlinear translations often challenge the linguistic norms of the target language, making their feasibility dependent on the structural similarity between the source and target languages. Despite these challenges, interlinear translation remains a valuable tool for those seeking to engage deeply with texts in their original form.

Despite the importance of interlinear translation, research on automating this process for ancient texts remains limited. To the best of our knowledge, our work is the first attempt to use machine translation to generate interlinear translations. We address this gap by evaluating state-of-the-art models for sequence-to-sequence problems in Ancient Greek on the task of interlinear translation from Ancient Greek to Polish and English. Our study compares the performance of general-purpose multilingual models with dedicated language models and assesses the impact of morphological tags and data preprocessing strategies on model performance.

Our research objectives include:

1. Evaluating the performance of state-of-the-art MT models in interlinear translation from Ancient Greek to Polish and English.
2. Comparing the effectiveness of general-purpose multilingual models with dedicated language models trained specifically on ancient languages and the given target language.
3. Assessing the impact of morphological tags on model performance by comparing different tag sets and various approaches to their integration.
4. Investigating the influence of pre-processing strategies, specifically focusing on normalization via removal of diacritics.

We focus on the full text of the Greek New Testament as our source corpus, considering its fundamental importance for international society, Ancient Greek being the original language, and the existence of numerous translations. For target languages, we examine differences in model performance with respect to languages with different syntactic characteristics – English (positional) and Polish (inflectional).
Our contributions are threefold:

1. Construction of a word-level-aligned parallel corpus of two interlinear translations of the Greek New Testament – to English and Polish, using data from Bible Hub and Oblubienica, respectively.
2. Fine-tuning experiments for interlinear translation using four base models – PhilTa, GreTa (Riemenschneider and Frank, 2023), and mT5 (Xue et al., 2020) (in two sizes), in 36 setups each, totaling 144 fine-tuned models
3. Novel approaches for encoding morphological information via dedicated embedding layers, which outperform solutions that do not utilize tags by up to 20% (BLEU score) on interlinear translation tasks into both target languages

# Methodology
## Datasets
### Data Acquisition

We prepared two corpora comprising interlinear translations of the Greek New Testament: one into Polish and one into English. The Polish dataset was scraped from Oblubienica, while the English dataset was obtained from Bible Hub. The corpora utilize different textual editions of the Greek text, with Oblubienica following NA28 and Bible Hub merging multiple textual variants.

### Corpus Alignment

To evaluate model performance based on the tag set used, we aligned the two corpora at the word level. We applied heuristics to match each word from one corpus with its counterpart in the other, successfully matching over 99% of the words. For unmatched words, we mapped morphological tags using the statistically most common counterpart or manual mapping for edge cases.

### Tag Sets

After alignment, each word in both corpora carries two morphological tags: one original and one cloned from the corresponding word-level counterpart in the other corpus. The tag sets vary in both quality and quantity, with Oblubienica having 1,073 unique tags and Bible Hub having 684.

## Text Preprocessing

We tested two preprocessing paradigms for Ancient Greek texts: one keeping all diacritics (using the spelling from Bible Hub) and another normalizing the data by stripping the texts of diacritics. This allows us to evaluate the impact of diacritics on model performance.

## Base Models

Our study employs four base models: GreTa, PhilTa, and mT5 in two sizes (base and large). GreTa and PhilTa are T5-base-sized models trained on Ancient Greek corpora, with PhilTa also including Latin and English in its training data. mT5 was pre-trained on the mC4 corpus, encompassing 101 languages including English and Polish, but not Ancient Greek. We selected mT5-base to match the size of the other models and mT5-large to explore the impact of increased parameters on performance.

## Model Inputs

We investigate five scenarios for incorporating morphological information, grouped into three categories:

1. Text Only (baseline): A scenario where no morphological information is provided.
2. Text With Morphological Tags (t-w-t): Morphological tags are encoded as part of the model's text input.
3. Embeddings (emb-*): Three scenarios utilizing a dedicated embedding layer for encoding morphological tags: a.

- Embeddings -- Sum (emb-sum): Morphological tag embeddings are summed with text embeddings.
- Embeddings -- Autoencoder (emb-auto): Tag embeddings are compressed, then decompressed and summed with text embeddings.
- Embeddings -- Concatenation (emb-concat): Compressed text and tag embeddings are concatenated.

## Training Details

Our dataset, comprising 7940 verses of the Greek New Testament, is split into three random subsets: training (95%, 7543 verses), validation (2.5%, 198 verses), and test (2.5%, 199 verses). We fine-tune 144 model combinations, varying target language, tag set, text preprocessing strategy, base model, and morphological information encoding method. To ensure fair comparisons, we trim each verse to the number of words that can be encoded by the least efficient setup among all parameter combinations.

Training employed an NVIDIA A100 GPU, with batch sizes adjusted based on memory constraints. For embedding-based approaches, we increased the learning rate for new neural network layers. We set a token limit of 256 across all scenarios.

Model performance is evaluated using BLEU scores (Papineni et al., 2002), with separator tokens removed from output sequences before evaluation to prevent rewarding models solely for correct output structuring.

# Results

Overall Performance:

- English translations generally outperformed Polish translations, but the top 40% of scenarios for both languages achieved similar BLEU scores (45-55).
- The close results suggest that the strict syntactical regime of interlinear translations may allow for cross-language comparisons, which are normally challenging due to differences in syntax and morphology.

Base Model Comparison:

- mT5-large outperformed other base models on average and achieved the best result in Polish translation.
- PhilTa secured the highest score for English translation, followed by GreTa and mT5-large.
- GreTa, despite not being pre-trained on English or Polish, performed similarly to mT5-base in both tasks.
- PhilTa struggled with Polish translation, failing to surpass a BLEU score of 30 in its best run.

Impact of Morphological Tags:

- Inclusion of morphological metadata consistently improved performance on the interlinear translation task, regardless of the chosen encoding strategy.
- The best-performing morphologically-enhanced models outperformed the baseline by approximately 20% for Polish (51.16 vs 42.65) and 21% for English (54.46 vs 44.95).
- Using a separate embedding representation for morphological information was preferable to passing it directly within the text.
- Among embedding-based strategies, sum-based methods (emb-sum and emb-auto) outperformed concatenation-based methods (emb-concat).

Tag Set Comparison:

- Both tested tag sets (Bible Hub and Oblubienica) yielded strong results, with Bible Hub's tag set slightly outperforming Oblubienica's in both translation tasks.
- The top-performing tag set (Bible Hub) had roughly 50% fewer forms, suggesting that either insufficient training data for less frequent forms in Oblubienica or differences in tagging quality may have influenced the results.

Impact of Preprocessing:

- In the vast majority of cases, regardless of the chosen tokenizer, runs with diacritics achieved better results both on average and in the best-case scenario.
- This finding challenges the common practice of removing diacritics in many experiments analyzing Ancient Greek texts.

# Conclusions

Our research demonstrates the effectiveness of neural approaches in interlinear translation of classic texts, specifically from Ancient Greek to English and Polish. The inclusion of morphological metadata significantly improves translation quality, challenging the notion that pre-trained transformers cannot utilize such information effectively in NLP tasks.

The superior performance of models pre-trained on both source and target languages (e.g., PhilTa for English) suggests that developing similar models for other language pairs could yield even better results. In the absence of such specialized models,

larger multilingual models or those pre-trained on the source language offer viable alternatives.
Our approaches to encoding morphological information, particularly the sum-based embedding methods, show promise for improving interlinear translation quality. These techniques could be applied to other low-resource language scenarios or tasks requiring precise structural preservation.
The benefit of maintaining diacritics in preprocessing highlights the importance of preserving nuanced linguistic information in MT tasks, especially for ancient languages. This finding may have implications for other NLP tasks involving Ancient Greek or similar languages with low resource settings and rich diacritical systems.

Bibliography

Gutt, E. A. (1991). Translation and relevance: Cognition and context. Blackwell.

Papineni, K., Roukos, S., Ward, T., & Zhu, W. J. (2002). Bleu: a method for automatic evaluation of machine translation. In Proceedings of the 40th annual meeting of the Association for Computational Linguistics (pp. 311-318).

Riemenschneider, A., & Frank, S. (2023). Exploring Ancient Greek with Language Models. Proceedings of the 3rd Workshop on Natural Language Processing for Digital Humanities.

Shuttleworth, M., & Cowie, M. (2014). Dictionary of translation studies. Manchester: St. Jerome Publishing.

Vaswani, A. (2017). Attention is all you need. Advances in Neural Information Processing Systems.

Xue, L., Constant, N., Roberts, A., Kale, M., Al-Rfou, R., Siddhant, A., Barua, A., & Raffel, C. (2020). mT5: A massively multilingual pre-trained text-to-text transformer. arXiv preprint arXiv:2010.11934.

BibleHub. Interlinear Bible. https://biblehub.com/interlinear/. Accessed: 2024-10-22.

Oblubienica. Ewangeliczny Przekład Interlinearny Biblii. https://biblia.oblubienica.eu/. Accessed: 2024-10-22.

Sanita Reinsone[1], Kyrre Kverndokk[2], Kati Kallio[3], Fredrik Skott[4], Mari Väina[5], Trausti Dagsson[6]

[1]University of Latvia, Latvia; [2]University of Bergen, Norway; [3]Finnish Literature Society, Finland; [4]Institute for Language and Folklore, Sweden; [5]Estonian Literary Museum, Estonia; [6]The Árni Magnússon Institute for Icelandic Studies, Iceland

Tradition Archives Meet Digital Humanities II

ID: 178

Half-day conference-themed workshop

Keywords: tradition archives, folklore collections, digital archiving, crowdsourcing, multilingual data

The workshop "Tradition Archives Meet Digital Humanities II" is proposed as a follow-up of the workshop with the same title that took place at the 2018 "Digital Humanities in the Nordic Countries" conference in Helsinki. This workshop seeks to provide an interdisciplinary discussion platform for researchers, archivists, and technology experts engaged in folklore archives, digital cultural heritage, and digital humanities, and interested in advancing computational folklore studies. The primary objective is to facilitate the exchange of insights, identify challenges, and promote innovative practices while mapping the current landscape of developments in this field.

Tradition or folklore archives are essential repositories that play a pivotal role in collecting, preserving, and studying intangible cultural heritage, particularly focusing on folklore and the broader field of cultural heritage. These archives have long been the custodians of vast collections of cultural expressions, including folk narratives, songs, rituals, oral histories, and other forms of intangible heritage. Their significance lies not only in the preservation of these materials but also in the systematic organization and classification that allow researchers to access, study, and analyze the diverse cultural expressions stored within them.

Aligned with the general thematic framework of DHNB2025, through an open call, the workshop will invite submissions that critically examine issues pertaining to community engagement in digital tradition archives, best practices for digital archive development, and the integration of advanced data analysis and AI-driven tools. The workshop will also bring attention to the complexities of linking multilingual folklore datasets, the application of computational analysis methods to folklore corpora, and the sustainability challenges in digital archives arising from the tension between project-based funding and long-term archival needs. The workshop seeks to enhance understanding of the current state of digital tradition archives and their impact on the field of (computational) folklore studies.

Importance of the workshop: As digital archives proliferate, they raise essential questions regarding community engagement, data integrity, and sustainability of digital resources developed. This workshop will provide participants with the opportunity to reflect on their experiences and develop strategies for collaboration and innovation in the field.

Target audience: the workshop is designed for a diverse audience, including:

- researchers in folklore studies and digital humanities;
- archivists and librarians involved in the management of cultural heritage collections;
- technology experts and developers working on tools for digital archiving;
- students and early-career scholars interested in folklore and digital humanities;
- community representatives engaged in preserving local traditions and narratives.

The anticipated number of participants is projected to be between 20 and 40. If the workshop could be organized in a hybrid format, it would likely attract a larger audience.

Expected outcomes: The workshop is expected to provide participants with up-to-date insights into recent developments in digital tradition archives and related projects while offering a platform for exploring new research ideas and fostering future collaborations. It will also strengthen collaboration within the professional network established in 2013 under the International Society for Ethnology and Folklore, specifically through the Working Group for Archives. By fostering continued dialogue on the intersection of folklore and digital humanities within this group, the workshop has the potential to lead to the creation of a new edition focused on digital folklore archives and computational folkloristics.

Format: The workshop will combine presentations and moderated roundtable discussions.

The ideal length of the workshop would be around 4–6 hours.

In terms of **technical requirements**, the workshop would require a projector for presentations and (if possible) audio playback equipment to facilitate any media or sound-related materials.

Anna Ristilä
University of Turku, Finland

Hot topics – Sentiment analysis on the plenary sessions of the Finnish parliament 1970-2020

ID: 154 / **Poster Session 2: 8**
Poster and demo (abstract) with accompanying a 1-minute lightning talk
Keywords: sentiment analysis, topics, parliament of Finland

The parliament of Finland, *eduskunta*, is a multi-party system with coalition governments and a generally consensus-driven political culture. More so than in a two-party system, for example, the many parties in *eduskunta* represent a wide spectrum of political views and reflect the views and attitudes present in the society. This paper will examine how the debates on “hard” and “soft” topics in *eduskunta* have evolved and become more or less adversarial between 1970 and 2020. There has been evidence of differentiation and general hardening of values in the Finnish society (e.g Helve 2023). The hypothesis for this paper is that sentiments towards "soft topics" such as *social and health care* and *development cooperation* have become less positive overall but also more adversarial, while "hard topics" such as *commerce* and *traffic and transport* have perhaps not become much more positive but less adversarial.

The data for this study contains Finnish plenary speeches from 1970 to 2020 (SEMPARL, Hyvönen et al. 2024). The speeches were previously modelled with a topic model (Ristilä & Elo 2023) which resulted in 26 topics that will be used as the categories for the sentiment analysis. The sentiment model that will be used has been specifically tuned for parliamentary speech (an XLM-R-parla model trained further with ParlaMint and EuroParl corpora).

Finnish parliamentary speeches have previously been studied from different angles, such as language identification (Jauhiainen *et al.* 2024) and topics (Ristilä & Elo 2023, Loukasmäki & Makkonen 2019), but not extensively with sentiment analysis (see however Tarkka *et al.* 2024, Proksch *et al.* 2019). We will use state-of-the-art sentiment analysis and examine which topics have been discussed with the most contradicting sentiments and how the levels of sentiment contradiction in certain topics have changed over time. This kind of study will provide a new perspective on what has been the most 'heated' topics in the Finnish parliament, and also how the societal attitudes have changed in Finland in the last fifty years.

Bibliography

Helve, Helena Marketta. 2023. ”Values and Solidarity of Young Finnish Millennials and Generation X”. Youth 3 (1): 401–13. https://doi.org/10.3390/youth3010027.

Hyvönen, Eero, Laura Sinikallio, Petri Leskinen, Senka Drobac, Rafael Leal, Matti La Mela, Jouni Tuominen, Henna Poikkimäki, ja Heikki Rantala. 2024. ”Publishing and Using Parliamentary Linked Data on the Semantic Web: ParliamentSampo System for Parliament of Finland”. Semantic Web. https://www.semantic-web-journal.net/system/files/swj3605.pdf.

Jauhiainen, Tommi, Jussi Piitulainen, Erik Axelson, Ute Dieckmann, Mietta Lennes, Jyrki Niemi, Jack Rueter, ja Krister Lindén. 2024. ”Investigating Multilinguality in the Plenary Sessions of the Parliament of Finland with Automatic Language Identification”. Teoksessa Proceedings of the IV Workshop on Creating, Analysing, and Increasing Accessibility of Parliamentary Corpora (ParlaCLARIN) @ LREC-COLING 2024, toimittanut Darja Fiser, Maria Eskevich, ja David Bordon, 48–56. Torino, Italia: ELRA and ICCL. https://aclanthology.org/2024.parlaclarin-1.8.

Loukasmäki, Petri, ja Kimmo Makkonen. 2019. ”Eduskunnan täysistunnon puheenaiheet 1999–2014: Miten käsitellä LDA-aihemalleja?” Politiikka 61 (2): 127–59.

Proksch, Sven-Oliver, Will Lowe, Jens Wäckerle, ja Stuart Soroka. 2019. ”Multilingual Sentiment Analysis: A New Approach to Measuring Conflict in Legislative Speeches”. Legislative Studies Quarterly 44 (1): 97–131. https://doi.org/10.1111/lsq.12218.

Ristilä, Anna, ja Kimmo Elo. 2023. ”Observing Political and Societal Changes in Finnish Parliamentary Speech Data, 1980–2010, with Topic Modelling”. Parliaments, Estates and Representation 43 (2): 149–76. https://doi.org/10.1080/02606755.2023.2213550.

Tarkka, Otto, Jaakko Koljonen, Markus Korhonen, Juuso Laine, Kristian Martiskainen, Kimmo Elo, ja Veronika Laippala. 2024. ”Automated Emotion Annotation of Finnish Parliamentary Speeches Using GPT-4”. Teoksessa Proceedings of the IV Workshop on Creating, Analysing, and Increasing Accessibility of Parliamentary Corpora (ParlaCLARIN) @ LREC-COLING 2024, toimittanut Darja Fiser, Maria Eskevich, ja David Bordon, 70–76. Torino, Italia: ELRA and ICCL. https://aclanthology.org/2024.parlaclarin-1.11.

C. Annemieke Romein[1,2,3]
[1]READ-COOP SCE, Austria; [2]Twente University, the Netherlands; [3]University Bern, Switzerland

Decoding the Past, Digitizing the Future: Transkribus and (Digitized) Cultural Heritage. An interactive tutorial

ID: 174
Full-day tutorial
Keywords: Automatic Text Recognition, Transkribus, Handwriting, Print, Layout Recognition, Digital Editions.

Tutorial Description

We cordially invite you to participate in an immersive investigation of Transkribus, a pioneering instrument at the nexus of digital humanities, archival science, and artificial intelligence. This hands-on workshop is aligned with DHNB 2025's overarching theme of "Digital Dreams and Practices." Transkribus is the ideal use case that integrates traditional humanities scholarship with computational techniques, brings the tools in an easily understandable way of academia to users with different backgrounds, and applies AI so that users benefit without learning how to program.

The Transkribus system was developed through EU-funded projects and employs advanced AI to automate handwritten text recognition, thereby reducing the time and expertise required for palaeographic analysis. With currently 225+ private and institutional members, 250,000 registered users, and 200+ public models available, Transkribus has significantly enhanced users' ability to adaptively tackle diverse handwriting styles, becoming an invaluable asset for researchers, archivists, and students.

The workshop is structured in two parts, allowing participants to engage with Transkribus at their own pace and following their interests and requirements. We encourage attendees to bring their own archival or library materials, whether handwritten or printed, reflecting the rich cultural heritage of the Baltic and Nordic regions. We will provide additional practice documents to ensure a comprehensive learning experience for those unable to bring their materials.

During the session, participants will receive practical experience using Transkribus' user-friendly, browser-based interface. They will learn how to navigate the platform, understand its AI-driven recognition process, and explore its potential applications in their research workflows. By the end of the workshop, participants will have acquired the skills to use Transkribus in their projects, thereby contributing to the digital preservation and accessibility of cultural heritage.

Aim of the Workshop

The principal objective of this workshop is to equip participants with the knowledge and expertise to utilize Transkribus proficiently and effectively within the context of their digital humanities research and/or archival practices. By the end of the session, attendees will:

1. Understand the principles behind AI-driven handwritten text recognition and its role in digital cultural heritage preservation.
2. Gain practical experience using Transkribus' interface, including uploading documents, training models, and extracting recognized text.
3. Learn strategies for integrating Transkribus into various research workflows, from individual projects to large-scale digitization efforts.
4. Explore Transkribus's potential to enhance accessibility and analysis of historical documents across different disciplines within the humanities.
5. Develop an appreciation for the intersection of traditional humanities scholarship, archival practices, and computational techniques in the digital age.

Practical details

Format:

- Interactive tutorial (or workshop) with practice moments for hands-on learning and practice.

Target Audience:

- Introductory Workshop: Suitable for those interested in learning about Automatic Text Recognition for the first time or first-time users of Transkribus.
- Advanced Workshop: For those who have attended the introductory session or already have experience using Transkribus. Interested participants without prior experience but who are keen to dive deeper are also welcome.

Anticipated Number of Participants:

- Flexible: Between 10 and 90 participants.

Technical Requirements:

- For the participants:
  - Internet connection.
  - A workspace with tables or desks to comfortably use a laptop.

Requirements for Participants:

- Device:
  - Bring your own laptop (preferred).
  - Tablets are possible but may offer reduced functionality.

    - A mouse can be helpful but is optional.
- Transkribus Account:
    - Create a free account at Transkribus.org before the workshop. This can also be done at the workshop, but pre-registration is more convenient.
    - Bring your registration details, including your email and password.
    - If you use multi-factor authentication (MFA), bring the device you use for authentication.

Bibliography

Muehlberger, Guenter, Louise Seaward, Melissa Terras, Sofia Ares Oliveira, Vicente Bosch, Maximilian Bryan, Sebastian Colutto, et al. (2019) ’Transforming Scholarship in the Archives through Handwritten Text Recognition: Transkribus as a Case Study’. Journal of Documentation 75, no. 5 (1 January 2019): 954–76. https://doi.org/10.1108/JD-07-2018-0114.

Terras, M., Anzinger, B., Gooding, P., Mühlberger, G., Nockels, J., Romein C. A., Stauder, A., Stauder, F. (2024). “The Artificial Intelligence Cooperative: READ-COOP, Transkribus, and the benefits of shared community infrastructure for Automated Text Recognition”. Submitted to Open Research Europe.

C. Annemieke Romein[1,2,3], Jos Mooijweer[4], Andreas Weber[1]
[1]University of Twente, Enschede, the Netherlands; [2]READ-COOP SCE, Innsbruck, Austria; [3]University of Bern, Switzerland; [4]Collectie Overijssel, Zwolle, the Netherlands

Digital Pathways to Regional Heritage: AI Solutions for Archive Access in Overijssel

ID: 316 / **Poster Session 2: 32**
Late-breaking poster
Keywords: Regional History; Handwriting Text Recognition; Visualising Archival Sources;

Archives play an indispensable role in safeguarding cultural heritage, ensuring the preservation of historical records for future research and analysis. The Overijssel Provincial States archive, which spans from 1578 to 1795, offers a distinctive insight into the region's historical governance and the nature of public petitions (1). The project employs Machine Learning/ Artificial Intelligence (ML/AI) to facilitate enhanced accessibility and comprehension of this invaluable resource, a fundamental anchor in the tapestry of regional history.

This project combines artificial intelligence-driven tools with human expertise in collaboration with several archival institutions, particularly the Collectie Overijssel. The objective is to render the archival content searchable and analyzable by 2028, with the dissemination of the results scheduled to coincide with the 500th anniversary of the Overijssel Provincial States in 2028.

This research integrates artificial intelligence innovations with human oversight, establishing a bridge between advanced computational techniques and cultural heritage studies. Citizen scientists' involvement is paramount, as they contribute to labelling datasets. The objective is to guarantee that the tools are accessible and effective for diverse audiences by fostering inclusive engagement.

In addition to its technical merits, this project highlights the broader significance of archives as repositories of collective memory. The Provincial States archive offers a particularly rich historical resource, featuring unique abstracts of requests—concise summaries that capture the direct petitions of local inhabitants to regional governance during the Republic’s tumultuous times. These abstracts represent what historians call "the voices of the people," providing an intimate glimpse into the everyday concerns and aspirations of citizens who sought intervention or support from provincial authorities (2). Each abstract is accompanied by a brief official decision, revealing whether the government rejected, approved, or acted upon the citizens' requests. This archival treasure trove is especially valuable because it has remained largely unexplored, with no existing index to facilitate access. By facilitating the search and interpretation of these historical records, we provide an invaluable resource for historians, educators, and the public, thereby fostering a deeper connection to Overijssel's rich administrative and social heritage (3).

Our approach demonstrates the powerful synergy between AI technology and humanities research, pioneering innovative methods for archival accessibility and historical interpretation. This poster illustrates the dynamic interplay between historical content, digital humanities techniques, and machine learning methodologies. By introducing a (sub)project focused on automated text recognition (ATR) of 60,000 pages (30,000 scans), we showcase how advanced visualization techniques—including heatmaps, geographical information, and chronological overviews—can comprehensively unlock and recontextualize archival materials.

The project's key milestones are as follows:

- In 2025, layout recognition and text recognition processes have been established, thus enabling the identification of document structures and implementing full-text search functionality.
- In 2026, the refinement of named entity recognition continues, supported by volunteer contributions to label and identify entities and metadata.
- In 2027, applying machine learning to topic modelling, clustering, and hierarchy development will facilitate the retrieval of topics with greater nuance, especially on requests. This entails utilizing tools such as ANNIF to augment the process of thematic analysis.

NWO (Dutch Scientific Organisation) funds this project:NWA.1518.22.105.

Bibliography

https://proxy.archieven.nl/20/7D277A2EA3FB493DA105ED573943FEDD

B. Waddell and J. Peacey, Introduction: power, processes and patterns in early modern petitioning. In: Waddell, Brodie and Peacey, J, (eds.) The Power of Petitioning in Early Modern Britain. (London, UK: UCL Press, 2024) pp. 1-32. https://eprints.bbk.ac.uk/id/eprint/53565/

E.D. Eijken et.al., In alle Staten; vierhonderd jaar provinciaal bestuur van Overijssel (Zwolle, 1978).

David Rosson
University of Helsinki, Finland

Reception Reader: intellectual history meets big data computing

ID: 105 / Poster Session 2: 6
Poster and demo (abstract) with accompanying a 1-minute lightning talk
Keywords: reception studies, text reuse, data mining, enlightenment, research infrastructure

The web tool “Reception Reader” provides a convenient visual interface for intellectual historians to study text reuse. The Computational History (COMHIS) research group in Helsinki has been working with the *Eighteenth Century Collections Online* (ECCO) dataset, which covers a major portion of all books published in the 1700s in Britain, for more than a decade, and has been making continuous improvements in metadata curation and digital infrastructure. This web tool showcases the latest developments built on top of these efforts.

As part of the *High Performance Computing for the Detection and Analysis of Historical Discourses* project (HPC-HD), we have applied a bioinformatics technique called BLAST over the entire ECCO text corpus, which proved to be very robust for detecting textual overlaps across more than 200k documents, even with noisy OCR data. This uncovered nearly a billion pairs of text reuse, namely, previously hidden links between these books.

While the COMHIS group is actively working with systematic approaches to answer historians’ research questions with data analysis, computer vision, natural language processing, and many fast-evolving digital humanities techniques, the Reception Reader interface aims to provide intuitive digital access to this vast amount of data for scholars from humanist traditions, and to expand their reach of the materials beyond what was previously conceivable with manual methods.

The user can start with one book of interest and see all the incoming influences from, and outbound connections to, other books. Which parts of the book were the most talked about, and during which periods? What did contemporary authors say about a particular section? What kind of debates and dialogues were ongoing, and what were the overall patterns? Reception, intertextuality, and authorship are just some examples of themes that can be explored with text reuse and this tool.

The nuances of meaning, and indeed findings for historians, may be discovered when text is understood in relation to other text, and between “distant reading” and “close reading”, between grasping the overall patterns and examining the specific context, between navigating an ocean of data and tracing each catch back to artefacts and evidence on a printed page. May this software be another step further along the journey of making sense of the burgeoning and spreading of ideas during the Enlightenment period.

Sasha Rudan[1,3], **Sinisha Rudan**[1], **Lazar Kovacevic**[1,2]
[1]LitTerra Foundation, Serbia; [2]Inverudio, USA; [3]Oslo University

ColaboFlow: A Platform for Visual and AI-Augmented Workflows in Digital Humanities

Thursday, 06/Mar/2025 9:00am - 9:30am
ID: 267 / Session LP 06: 1
Long paper (abstract) | 20-minute presentation with a 10-minute Q&A
Keywords: AI-Augmented Workflows, Digital Humanities, Collaborative Research, Visual Workflow Management, Reproducibility in Research

ColaboFlow and LitTerrra

ColaboFlow is a framework integrated into the LitTerra platform, designed to enable researchers in the digital humanities to explore, transform, and visualize *digital editions and multilingual corpora* using **visual and AI-augmented workflows**. This paper presents ColaboFlow as a flexible infrastructure supporting the creation, execution, and sharing of workflows that are both user-friendly and scalable. The platform offers a comprehensive toolset, bridging the gap between researchers **with no programming expertise** and those requiring **advanced, automated** workflows.

ColaboFlow allows for the **visual creation** of workflows through *BPMN (Business Process Model and Notation) diagrams*, making it accessible to users with varying technical backgrounds. Researchers can define workflows through an intuitive interface and interact with them directly, facilitating a seamless translation of academic inquiry into actionable processes. . Through this combination of accessibility and computational power, ColaboFlow enables researchers to engage with large and complex datasets in innovative ways.

The paper discusses **semantic capabilities** of ColaboFlow, both workflows and the datasets they manipulate. These semantic descriptions ensure that the system, including the AI components, can comprehend, interact with, evolve (transform) and execute workflows accurately. This semantic infrastructure is essential for the platform's ability to support complex and reproducible workflows, that should handle multilingual corpora, conduct cross-lingual analyses, and integrate various forms of multimedia enrichment, external service extensions of workflows or datasets and present it in a form of a coherent research process.

Workflow Solidification Process

A key feature of ColaboFlow that we are investigating through this work is the concept of ***workflow solidification***, a process where descriptive workflows are transformed into task-structured and ultimately fully executable workflows. In such a way, a loose workflow is *"tightened up"* into a solid workflow - hence *workflow solidification*. This process occurs in three stages. In the first stage, (i) researchers describe their workflows at a high level, focusing on objectives and broad tasks. In the second stage, (ii) these descriptions are transformed into *"taskative" workflows*, which define specific tasks and their relationships. The final stage involves (iii) converting these workflows into executable models that can be applied to real-world datasets.

The workflow solidification process not only ensures that researchers have a *clear and structured plan* for their analyses, but also provides a pathway toward *automation and scalability*. Researchers can visualize their workflows, refine them through iterative testing, and modify tasks as needed. This process facilitates a **deeper engagement** with both the research questions and the datasets, enabling scholars to focus on intellectual challenges rather than technical execution.

Visual and AI-Augmented Interaction

ColaboFlow integrates a powerful visualization component, allowing researchers to interact with their workflows' results using AI-augmented tools. Visualization is essential in helping researchers interpret large-scale data and workflows, enabling them to refine their research work and analyses in response to emerging patterns. The framework employs a structured approach to visualization based on ***"The Grammar of Graphics"*** (Wilkinson, 2012), where users can generate charts and graphs as part of their workflow process.

The interaction between workflow execution and visualization creates a feedback loop in which researchers continuously refine their workflows based on the visual output. This dynamic interaction is particularly valuable in complex or large datasets, where distant-reading itself may not reveal significant insights. By making the syntetised data interactive and accessible through AI-driven visualization, ColaboFlow empowers researchers to make more informed decisions throughout their research processes.

Collaborative and Reproducible Research

One of the central goals of ColaboFlow is to facilitate **collaborative research** across diverse teams and institutions. The platform is designed with **reusability and reproducibility** in mind, ensuring that workflows can be **shared, modified, and executed** across different research contexts. Each workflow is **versioned** , allowing for specialization to particular research tasks or corpora. Researchers can modify and extend workflows based on specific needs, while maintaining the integrity of the original design through *detailed provenance tracking*.

The **provenance tracking** is essential in digital humanities research, particularly when working with multilingual or cross-cultural corpora. By enabling version control and documenting the evolution of workflows, ColaboFlow ensures that research remains **transparent** and **verifiable**. Moreover, the platform's collaborative infrastructure allows multiple researchers to work together on the same **project**, sharing their workflows and datasets, and building upon each other's work.

Human-on-the-Loop Design and Focus Group Feedback

The design of ColaboFlow incorporates a ***human-on-the-loop*** approach, where the platform's AI capabilities augments and assist researchers while keeping them central to the decision-making process. This design philosophy emphasizes that while AI can automate many aspects of research, human input remains crucial in guiding workflow composition, modification, and interpretation. By incorporating AI augmentation in a manner that supports rather than replaces **human judgment and decision-making**, ColaboFlow democratizes DH reasearch and ensures that the platform remains accessible to researchers of all technical backgrounds.

To evaluate the design and usability of ColaboFlow, a series of focus groups were conducted with practitionaries and researchers from different fields.

Application of ColaboFlow: A Demonstration

While this paper focuses on the design and infrastructure of ColaboFlow, it is important to illustrate its application in real-world research scenarios. The paper briefly demonstrates ColaboFlow's capabilities using three multilingual corpora as examples: **Henrik Ibsen**'s multilingual corpus from the ***Centre for Ibsen Studies***, **Jane Eyre**'s multilingual corpus from the ***Prismatic Jane Eyre project***, and **Vladimir Nabokov**'s multilingual work from the ***IMPULZ project***. These examples show how researchers can use ColaboFlow to structure workflows that analyze translations, conduct comparative textual analyses, and visualize linguistic patterns across different languages.

Though the focus of this paper is on ColaboFlow as an infrastructure, these examples demonstrate how the platform supports complex, multilingual text analysis by offering scalable, reproducible, and AI-augmented workflows. The flexibility of ColaboFlow's infrastructure allows for the handling of diverse research tasks, from simple text annotation to complex cross-linguistic comparisons.

Conclusion and Future Directions

ColaboFlow is a comprehensive platform designed to address the needs of digital humanities researchers through its visual and AI-augmented workflows. By offering a user-friendly interface that supports both novice and advanced users, the platform bridges the gap between descriptive and fully executable workflows. The preliminary results from focus group research demonstrate the platform's efficacy in facilitating collaborative, reproducible, and scalable research.

Moving forward, ColaboFlow's development will focus on enhancing its AI capabilities, streamlining workflow execution, and expanding its visualization tools. These future developments aim to make the platform more accessible and empowering, ensuring that it continues to meet the evolving needs of digital humanities researchers while at the same time keeping aligned with AI-ethics and human-on-the-loop principles that are becoming crucial more than ever.

Sasha Rudan[1,4], **Sinisha Rudan**[1], **Lazar Kovacevic**[3,1], **Eugenia Kelbert**[2,1], **Lucija Mandic**[2]
[1]LitTerra Foundation, Serbia; [2]Institute of World Literature, SAS, Slovakia; [3]Inverudio, USA; [4]Oslo University

Exploring, transforming and visualizing digital editions and corpora with visual and AI augmented workflows

ID: 263
Half-day conference-themed workshop
Keywords: AI-Augmented Workflows, Visual Workflow Solidification, Digital Humanities, Multilingual Corpora, Collaborative Research

At the workshop we will use ColaboFlow framework integrated in the LitTerra platform for presenting multilingual corpora to explore and transform digital editions and corpora with visual and AI augmented workflows.

We will provide 3 corpora for participants to work with: 1. Henrik Ibsen's multilingual corpus from the Centre for Ibsen Studies at the University of Oslo 2. Jane Eyre's multilingual corpus from the Oxford University project, Prismatic Jane Eyre 3. Vladimir Nabokov's multilingual work from the IMPULZ project IM-2022-68

The ***ColaboFlow framework*** is an expressive tool for creating, executing, and sharing visual and AI augmented workflows. On one hand, it is designed to be user-friendly and accessible to researchers with no programming skills. Therefore it supports BPMN diagrams as a visual representation of workflows and interaction with them. On the other hand, it is designed to support fully executable workflows that can be executed in a scalable and distributed environment. It provides semantic description of datasets propagating through workflows and semantic description of the workflows themselves. This enables AI to correctly comprehend workflows that can be composed, extended and executed.

During the workshop, we will have hands-on session where participants will be able to practice with such a process. We will practice the process we call ***workflow solidification*** where participants will start with describing their own workflows, then solidify them from descriptive to "*taskative*" (task-structured), and finally to fully executable workflows. The purpose of these steps is to end up with workflows that can be executed on real corpora and datasets in order to generate research-required results. During that process participants will be able to visualize their workflows, refine them, expand them, and interpret the results.

After the solidification process, and executing the workflows on real data, participants will be able to **visualize results** using AI augmented charts and graphs generation feature of LitTerra platform. The visualization process will follow the "***The grammar of graphics***" (Wilkinson 2012) model, visualized as a workflow, where participants will be able to interact with the visualizations and modify them. This will enable them to interpret the results and to make further decisions on the research process.

Finally, the will be able to use the LitTerra platform to augment the corpora with both visualizations and distant reading augmentation (such as showing the charts along the text and annotating the text with the annotations generated from the workflow results). This will enable them to explore the corpora in a new way and to make new discoveries which will eventually start a new cycle of the research process.

ColaboFlow workflows are designed to be collaborative and reusable/reproducible. This means that we will practice collaboration across research teams and annotation of the workflows, as well as reusing the workflows and corpora in the next research cycles. To be reproducible, the workflows support versioning, specialization (for specific parts of corpora, like specific tools for specific text languages), and provenance tracking, while datasets support versioning and provenance tracking. Having these features, we would be able to run the identical workflows against different text witnesses or their translations aiming for solid and argumented comparative results.

The setup of the workshop will be as follows: - Introduction: We will start with a brief introduction to the ColaboFlow framework and the LitTerra platform for participants to get familiar with the concepts and the benefits of the tools and overall methodology.

Potential and challenges: After that we will have a group discussion on the potential use cases and the challenges that participants are facing in their research. - Human-in-the-loop: We will discuss the concept of human-in-the-loop particularly in the context of AI capable of composing and executing research workflows. - Hands-on session: The main part of the workshop will be a hands-on session where participants will be able to explore and transform digital editions and corpora with visual and AI augmented workflows. We will provide a set of predefined workflow descriptions and corpora for participants to work with. Participants will be able to modify the workflows and corpora and see the outcome results. - Evaluation: We will evaluate participants' experience and the resulting workflows and their results. We will present some of them separately and provide overall statistics in behavioral patterns and outcomes. - Discussion: We will conclude the workshop with a discussion on the potential future developments and the benefits of using visual and AI augmented workflows in digital humanities research. We will also discuss the potential future collaborations and the ways to continue the work started at the workshop.

Bibliography

Wilkinson, Leland. The grammar of graphics. Springer Berlin Heidelberg, 2012.

Rudan, Sasha Mile, Sinisa Rudan, and Birger Møller-Pedersen. "Extending BPM (N) to Support Face-to-Virtual (F2V) Process Modeling." MODELSWARD. 2021.

Rudan, Sasha Mile, et al. "Twin Talk: Bukvik+ LitTerra+ Colabo. Space-An example of DH collaboration across disciplines, languages, and style." (2020): 15-29.

Rudan, Sasha Mile, et al. "Colabo. Space-Participatory Platform for Evolving Research and Publishing Workflows." Linking Theory and Practice of Digital Libraries: 25th International Conference on Theory and Practice of Digital Libraries, TPDL 2021, Virtual Event, September 13–17, 2021, Proceedings 25. Springer International Publishing, 2021.

Reynolds, Matthew, et al. Prismatic Jane Eyre: Close-Reading a World Novel Across Languages. Open Book Publishers, 2023.
Reynolds, Matthew, and Giovanni Pietro Vitali. "Mapping and Reading a World of Translations: Prismatic Jane Eyre." Modern Languages Open 1 (2021).

Rothwell, Andrew, Andy Way, and Roy Youdale, eds. Computer-Assisted Literary Translation. Taylor & Francis Group, 2023.

Sasha Rudan, **Sinisha Rudan**
ChaOS, NGO, Serbia

CoLaboArthon - A Bridge Between Socially Engaged Art and Science for Collective Impact

ID: 273 / Poster Session 2: 29
Long paper (full-text) | 20-minute presentation with a 10-minute Q&A
Keywords: Socially Engaged Art, Collaborative Creativity, Digital Humanities, Art-Science Dialogue, Interdisciplinary Collaboration

CoLaboArthon is an interdisciplinary framework that integrates art, technology, and science to foster collective creativity and address societal challenges. Designed by ChaOS and partners, this initiative merges socially engaged art and digital tools to create spaces for dialogue and collaboration. Through participatory methodologies and the CoLabo.Space ecosystem, CoLaboArthon enables artists, scientists, and community members to co-create solutions to pressing social issues like migration, sustainability, and disability.

The CoLaboArthon process engages participants in storytelling, role-playing, and collective art-making to address these complex topics. Workshops culminate in interactive performances, where participants and audiences alike contribute to shaping the final creative outcome. One prominent example is "Poetry on the Road," an international dialogue where over 50 poets collaborated in multiple languages on themes of migration and human rights, transcending cultural boundaries through art.

CoLaboArthon also emphasizes the use of technology in facilitating collaboration. The CoLabo.Space platform integrates machine learning and knowledge mapping to enhance the creative process, clustering participants based on shared goals and ideas. In the "Everyday Heroes" project, participants used AI-driven tools to visualize collective aspirations for a better world while addressing the United Nations' Sustainable Development Goals (SDGs).

CoLaboArthon offers a dynamic model for interdisciplinary collaboration, combining art, science, and technology to inspire social change. Its innovative approach holds great potential for advancing digital humanities and socially engaged art through collective action and creativity.

Liina Saarlo[1], **Kati Kallio**[2,3], **Maciej Janicki**[3], **Mari Väina**[1]
[1]Estonian Literary Museum, Estonia; [2]Finnish Literature Society, Finland; [3]University of Helsinki, Finland

Copies, duplicates, forgeries: surprises via similarity calculation

ID: 313 / WS10: 8
Tradition Archives Meet Digital Humanities II
Keywords: Oral tradition, similarity recognition, duplicates, literary influence, authenticity

Reliability, authenticity, and uniqueness are some of the most arduous and challenging issues regarding the historical documents of oral tradition in folklore archives. Here, authenticity refers to the character of different archival documents in relation to the historical oral traditions they are supposed to represent, while it has also been analysed as a scholarly ideology with cultural and political implications (Bendix 1997).

The archival collections may include copies, edited versions, forgeries, and write-offs that are laborious to identify. Manual text-critical processing of large corpora would often require superhuman knowledge of the details of the corpora, use of extensive reference materials, encyclopaedic knowledge of literary and cultural history, and a considerable investment of time. Often the fruits of this kind of work have been frustration and suspicion, blaming universal literacy for the destruction of the archaic culture, and sometimes leading to the downfall of intellectual or academic heroes into the dust of plagiarism.

Computational identification of textual similarities does not automatically solve these kinds of problems but does allow the identification of relevant cases for further examination, especially when supplemented with analysis of metadata such as place and time of recording. Variations with simultaneous stereotypical or formulaic expressions are inherent in oral tradition (e.g. Foley 1995; Harvilahti 1992). Texts that are too similar or too unique often indicate non-oral or non-traditional origins. In digitised corpora, the computational similarity analysis may facilitate the identification of texts that are excessively similar or unique, likely due to factors such as individual creation, copying, falsification, or literary origin. This allows researchers to test hypotheses about the character of archival texts and their collectors.

Rather than being seen as (only) problems or issues, or as a reason to create unofficial "black lists" of songs, performers, collectors or publishers to be ignored in serious research, such incidents provide interesting insights into, for example, creativity in the oral tradition, complex oral-literary relationships (e.g. DuBois1996), networks of professional and amateur collectors (e.g. Mikkola 2013), different aims and ambitions of singers and recorders (e.g. Sarv 2012), and archival curation processes in different historical periods (e.g. Harvilahti et al. 2018).

In our paper we present what we have discovered exploring the large oral poetry corpora of Finnic runosong tradition with the set of similarity detection methods developed in the course of the FILTER project (2020–2024; see e.g. Janiki et al. 2024; 2023). We invite participants to think along and discuss about the spectrum between uniqueness and full or close similarity within the tradition and archival collections, and how computational tools and methods enhance the understanding of the folkloric versus literary creativity.

Bibliography

Bendix, Regina 1997: In search of authenticity: the formation of folklore studies. Madison (WI): University of Wisconsin Press.

DuBois, Thomas A. 1996. The Kalevala Received: From Printed Text to Oral Performance. Oral Tradition, 11/2, 270–300.

Foley, John Miles 1995. The Singer of the Tales in Performance. Bloomington and Indianapolis: Indiana University Press.

Harvilahti, Lauri 1992. The Production of Finnish Epic Poetry – Fixed Wholes or Creative Compositions? Oral Tradition, 7(1), 87–101.

Harvilahti, Lauri, Audun Kjus, Clíona O'Carroll, Susanne Österlund-Pötzsch, Fredrik Skott and Rita Treija (eds.) 2018. Visions and Traditions. Knowledge Production and Tradition Archives. Helsinki: Academia Scientiarum Fennica

Janicki, Maciej, Eetu Mäkelä, Mari Väina ja 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.

Janicki, Maciej & Kati Kallio & Mari Sarv 2023. Exploring Finnic written oral folk poetry through string similarity. Digital Scholarship in the Humanities 38 (1): 180–194. https://doi.org/10.1093/llc/fqac034

Mikkola, Kati. 2013. "Self-taught collectors of folklore and their challenge to archival authority." White Field, Black Seeds: Nordic Literacy Practices in the Long Nineteenth Century. Helsinki: SKS, pp. 146-157.

Sarv, Mari (ed.). 2012. Regilaulu müüdid ja ideoloogiad. Tartu: EKM Teaduskirjastus.

Werner Scheltjens
University of Bamberg, Germany

STRO 2.0 - Re-engineering Sound Toll Registers Online

Wednesday, 05/Mar/2025 2:30pm - 3:00pm
ID: 108 / Session LP 03: 3
Long paper (full-text) | 20-minute presentation with a 10-minute Q&A

Keywords: Sound Toll Registers Online, data modeling, fiscal history, re-engineering, refactoring

The database Sound Toll Registers Online (short: STRO) contains millions of records that describe the commodities carried on ships passing through the Danish Sound between 1497 and 1857. STRO is based on a taxation register kept on behalf of the Danish Crown. This register has enjoyed significant historical interest as a unique source for early modern trade and shipping in Northern Europe. This traditional interest has also guided the digitization project. The current data model of STRO focuses on ‘passages’ and ‘cargoes’, but does not fully capture the fiscal nature of the source. This hampers our understanding of the data in STRO. Most importantly, the current model offers little support for analysing STRO from the perspective of fiscal history. The paper argues that re-engineering STRO provides a first step towards overcoming this limitation, and describes how it is done. First, the data model is restructured around the concept of 'customs entries' to underline the fiscal nature of the source and new entities are created to normalize the model. Second, parts of the data schema pertaining to tax data are refactored to make the model more efficient and maintainable. These changes drastically reduce the redundancy and sparsity of the current database. While targeting key components of the original data model, STRO 2.0 simplifies data handling and support, paves the way for further normalization of the new data model and opens new avenues for innovative historical analysis.

Raivis Sīmansons, **Cory McLeod**
Žanis Lipke Memorial

Žanis: Through Our Eyes. Multi-Plot Documentary for Virtual Reality. Presentation and discussion about digital Holocaust memory and immersive learning

ID: 287 / WS02: 6
Innovations and New Interactions through Digital Cultural Heritage in GLAM Sector
Keywords: digital cultural heritage, innovations, co-creation, audience engagement, new forms of interaction, immersive learning

Considering limitations in Holocaust learning posed by the post-witness era, the archival material and physical space where historic events took place proves to be the last authentication factor of Nazi crimes.

The story of Žanis is a mosaic of the 55 people he rescued. Each person's experience was different, and each rescue was its own uniquely harrowing endeavor. We learn the story of Lipke through the individual stories of the people he rescued and through historical sites where events took place. Each story is told through a first-person narrative, using original written and oral sources, and digital reconstruction of historical sites in 3D.

These stories are inherently non-linear, with intertwining timelines, overlapping characters, and different points of view. Virtual Reality as a medium is also inherently non-linear, lending itself seamlessly to an immersive, choose-your-own-adventure narrative format.

HANNA'S STORY is the first in this series. It recounts the experiences of Hanna Stern, who, at the age of six, was deported from Berlin to the Riga Ghetto with her mother, Sophie, and her older brother, Phillip. While interned at the Kaiserwald Concentration Camp, Hanna was saved by Žanis, who helped her escape and hid her until the end of the Holocaust. This story is

based on a letter Hanna's mother, Sophie, wrote to a lawyer representing their petition to the government of the Federal Republic of Germany for wartime compensation. The voiceover is read by Hanna's daughter, Ilana Avimor.

The expected outcome of presentation and discussion with digital humanities experts are eventual results from an impromptu front-end evaluation of a work in progress of this new VR production which features an experimental out-of-headset mode of presentation suitable for a group experience.

https://zanisvrdoc.com/

Bibliography

Daniela, L. (ed.). (2020). New Perspectives on Virtual and Augmented Reality. Finding new ways to teach in a transformed learning environment. Routledge.
Verschure, P. (2021). Digital Holocaust Memorialisation. Digital Holocaust Memory, 2021. https://reframe.sussex.ac.uk/digitalholocaustmemory/online-discussions/

Verschure, P. M. J., Wierenga, S. (2021). Future memory: a digital humanities approach for the preservation and presentation of the history of the Holocaust and Nazi crimes. In: Holocaust Studies: A Journal of Culture and History. https://doi.org/10.1080/17504902.2021.1979178

Walden, V. G. (ed.). (2022). The Memorial Museum in the Digital Age. REFRAME Books. https://reframe.sussex.ac.uk/the-memorial-museum-in-the-digital-age/

Maria Skeppstedt[1], Magnus Ahltorp[2], Gijs Aangenendt[1,3], Ylva Söderfeldt[3]

[1]Centre for Digital Humanities and Social Sciences Uppsala, Department of ALM, Uppsala University; [2]Language Council of Sweden, Institute for Language and Folklore; [3]Department of History of Science and Ideas, Uppsala University

Further developing the Word Rain text visualisation technique in a digital history project

Wednesday, 05/Mar/2025 1:30pm - 2:00pm

ID: 244 / Session LP 01: 1

Long paper (full-text) | 20-minute presentation with a 10-minute Q&A

Keywords: Text visualisation, Digital history, Word embeddings

The development of the Word Rain text visualisation technique started as a theoretical project, in which insights from previous NLP and text visualisation research were used for creating a visualisation that addressed some of the problems associated with the classic word cloud. In particular, the problem of the word clouds being unsuitable for the analytical tasks to which they are often applied within digital humanities — such as categorising words and comparing texts — was addressed. The next step in the development of the Word Rain visualisation technique consisted of applying the technique to real use cases. One of these use cases studied longitudinal content change in periodicals from diabetes patient organisations. This work, which was carried out in a (partly digital) history project, resulted in several features being added to the Word Rain visualisation, e.g., features related to word positioning, colour use and word prominence calculations. We here describe this work of practically evaluating and further developing the Word Rain text visualisation technique, as well as present and motivate the new features of the visualisation technique that the work resulted in.

Asta Skujytė-Razmienė
Institute of Lithuanian Literature and Folklore, Lithuania

The Does and Don’ts of Volunteer-based Digitalisation Initiative at the Lithuanian Folklore Archives

ID: 297 / WS10: 1
Tradition Archives Meet Digital Humanities II
Keywords: Folklore, digitalisation, volunteer, community, archives

This presentation focuses on methods and challenges in engaging communities in the development of digital archives, emphasizing the role of volunteering in document digitization at the Lithuanian Folklore Archives.

Without a doubt, involving community in the process of preserving cultural heritage strengthens the bond between interested participants and their heritage, fosters the sense of inclusivity, and, last but not least, helps to promote the institution and its work. However, effective participation requires well-planned strategies that align both with the technical demands of digitization and with the capabilities (and personal motivations) of those who apply to volunteer.

The key method is one on one meetings and training sessions between archivists and volunteers, as well as working with user-friendly digitization tools. Nevertheless, maintaining consistent engagement is challenging, as well as ensuring data quality, and addressing the digital divide. We noticed that volunteers vary in technical expertise, requiring resource-intensive oversight, however, there is little room for the feedback, as most of the volunteers work just for a very brief period of time. Additionally, intellectual property concerns and responsibility for the equipment involved in the process of digitalization necessitate guidelines to safeguard both the Archive’s and community’s interests.

I hope that by addressing these challenges, the physical involvement of the community members in the processes of archival digitalisation can become a sustainable model that not only helps to preserve cultural heritage, but also empowers people to explore their own connection with traditions and folklore.

Steinþór Steingrímsson
The Árni Magnússon Institute for Icelandic Studies, Iceland

Disseminating News using Machine Translation: An MT Researchers' Perspective

Friday, 07/Mar/2025 2:00pm - 2:20pm
ID: 228 / Session LP/SP 10: 2
Short paper (abstract) | 15-minute presentation with a 5-minute Q&A
Keywords: Machine Translation, evaluation, AI, AI appropriateness

Machine translation (MT) has reached the stage where the technology can produce fluent texts in dozens if not hundreds of languages. What does that mean for how we use the technology? MT usage is sometimes divided into two categories: (i) gisting, for when the user needs to get an impression of the meaning of a text they do not understand. In this case an imperfect translation is often helpful. (ii) dissemination, in which MT is a step in production of a text that will be published (see e.g. Moorkens et al. 2024:153). When using MT to produce publishable texts, we can select between a number of approaches, the most common being unedited MT output and post-edited or post-corrected output. In the latter case humans edit a machine-generated translation to an acceptable form. While post-editing usually results in more acceptable translations than unedited output, the efficacy is debated. The most important thing should be to convey the message clearly and effectively to readers. If using MT in any way fails to do that, its use should be reconsidered. Santy et al. (2021) suggest an interactive approach, using MT to suggest translations to translators as they are typing. They run an experiment where they compare translations without the use of MT, translations using post-editing and translations using interactive MT. They find that using interactive MT produces translations of similar or better quality than translations that don't use MT, in less time, while the post-edited translations are of less quality than the other two categories.

European Perspective (EP) is an initiative by 20 European news broadcasters that aims to ``give citizens a window into the issues affecting Europe locally and globally; bringing audiences quality journalism in their own language.'' To reach this aim, content from the participating organizations is translated to other European languages using AI (see https://www.europeanperspective.net/home).
Accompanying the translated stories is a disclaimer. It varies slightly between languages, but delivers the message that the text is ``Translated using Artificial Intelligence'' and in some cases the MT engine used is named. Sometimes the web pages showing the translated news stories contain a link to the original text, but sometimes they do not.

I investigate news published by RÚV, the Icelandic National Broadcasting Service, within the European Perspective project in the months of August and September 2024. While the news agency announced, when launching their participation in the project, that all MT-generated text would be checked and edited, I find that almost half the texts have minimal changes and that the average quality of their texts is measured lower than the average quality of texts in other foreign news stories.

While no reference translations made by human translators are available, I try to evaluate whether the edited MT translations published by RÚV within the European Perspective project are of less quality than human translated text.

Following the evaluations I discuss the possible effects this has on readers, whether using AI to generate texts for publishing may be a risk for the news agency and how such possible risks may be averted. I discuss the choice of a translation service, how it should be used and why it is important to make an informed decision when such a service is selected.

Furthermore, I discuss the need for an MT label (see discussion in e.g. Simard, 2024). Within the EP project there is no standardized label for indicating how the text was generated. I argue the case for it being a missed opportunity for the news agencies, as improper use of AI can lead to people losing trust in what they publish. It has been argued that all MT and AI generated content, which has not been changed to such an extent that a person can claim full responsibility for it, should be labelled. I set forward suggestions for how that could be carried out in order to make MT generated text more publishable.

Bibliography

Joss Moorkens, Andy Way, Seamus Lankford. 2024. Automating Translation. Routledge.

Sebastin Santy, Kalika Bali, Monojit Choudhury, Sandipan Dandapat, Tanuja Ganu, Anurag Shukla, Jahanvi Shah, and Vivek Seshadri. 2021. Language Translation as a Socio-Technical System:Case-Studies of Mixed-Initiative Interactions. In Proceedings of the 4th ACM SIGCAS Conference on Computing and Sustainable Societies (COMPASS '21). Association for Computing Machinery, New York, NY, USA, 156–172. https://doi.org/10.1145/3460112.3471954

Michel Simard. 2024. Position Paper: Should Machine Translation be Labelled as AI-Generated Content? In Proceedings of the 16th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Track), pages 119–129, Chicago, USA. Association for Machine Translation in the Americas.

Steinþór Steingrímsson, Bjarki Ármannsson
The Árni Magnússon Institute for Icelandic Studies, Iceland

In Pursuit of the Trivial

Friday, 07/Mar/2025 2:40pm - 3:00pm
ID: 233 / Session LP/SP 10: 4
Short paper (full-text) | 15-minute presentation with a 5-minute Q&A
Keywords: Large Language Models, Icelandic, Automatic Question Answering

We compare the performance of state-of-the-art Large Language Models on a recently released bench- marking set for automated question answering for Icelandic and compare it with performance on questions from an Icelandic trivia game. We find that the models perform worse for questions on Icelandic subjects, specifically Icelandic culture, but somewhat surprisingly do better on a trivia game for people than on the benchmark set meant for language models, built around data that the model has seen during training. We also call into question some aspects of the benchmarking set and discuss what playing trivia games can tell us - if anything - about the capabilities of these models.

Eva Strazdina, Zanda Rubene
University of Latvia, Latvia

Implementation of Films as Innovative Educational Tools in Latvia (1920s-1930s)

ID: 181 / Poster Session 2: 14
Poster and demo (abstract) with accompanying a 1-minute lightning talk
Keywords: education innovation, educational film, source analysis, Digital Library of Latvia, Latvian State Historical archive

Technological developments at the beginning of the 20th Century influenced work in the various public areas, including education. This research explores the implementation of films as innovative educational tool in Latvia (1920s-1930s). At the time media education was developed and implemented in the educational process in accordance with similar initiatives in other European countries. Many countries, including Latvia, introduced the concept of Kulturfilm ("cultural film" or "educational film"). To provide a modern education in Europe, teachers had to acquire the necessary technical skills, the psychological foundations of film use and didactics. The guiding research questions are: what were the characterizing factors of the film inclusion in the educational process and which childhood discourse in the inclusion of films in education was dominant in Latvia in the 1920s and 1930s? The research method is a source analysis, which includes the source criticism and interpretation stages. The implementation of films as innovative educational tools is discussed in relation to the culture, leading ideas and ideology of the era by analyzing two different written source groups including periodicals selected through the Digital Library of Latvia and policy documents currently stored at the Latvian State Historical archive. For source interpretation, the key categories were defined: 1) the institutional approach; 2) foreign experience; 3) film as an educational tool; 4) film demonstration outside of schools; 5) provision of projectors and film demonstration in schools. As a result, the work reports the main aspects of the implementation of educational films as teaching tools promoting innovative teaching methods in Latvia in the 1920s and 1930s by analyzing those through the dominating childhood discourse at that time.

Bibliography

Andersons, E. (1982) Latvijas Vēsture 1920–1940. Stockholm: Daugava/Universaltryck Grafiska AB

Bühler-Niederberger, D. (2011) Lebensphase Kindheit. Theoretische Ansätze, Akteure und Handlungsräume. Weinheim, Basel: Juventa.

Buckingham, D. (2005) New media, new childhoods? Children's changing cultural environment in the age of digital technology. In: Kehily, M., J. (Ed.) An Introduction to Childhood Studies. Berkshire, Open University Press, pp. 108 -122.

Buckingham, D. (2010) Media Education. Literacy, Learning and Contemporary Culture. Cambridge, Polity.

Fuchs, E., Bruch A., Annegarn-Gläß, M.(2016) Educational Films: A Historical Review of Media Innovations in Schools. Journal of Educational Media, Memory, and Society, 8 (1), doi: 10.3167/jemms.2016.080101

Hüther, J., Podehl, B. (2005) Geschichte der Medienpädagogik. In: Hüther, J., Sch

Key, E. (2012) Das Jahrhundert des Kindes. Weinheim, Basel, Beltz Verlag.

Krūmiņš G., Zemītis G., Strenga G., Straube G. Šiliņš J., (2016) Ieskats Latvijas vēstures svarīgākajos jautājumos. Rīga, Valsts Kanceleja

Ķestere, I. (2005) Pedagoģijas vēsture. Rīga, Zvaigzne ABC.

Penģerots, V. (1927) Kinematografs kā mācības un audzināšanas līdzeklis. No: Izglītības ministrijas mēnešraksti T. Zeiferta redakcijā. II izdevums. Rīga: Izglītības ministrija, 1927, 19.-25.lpp.

Postman, N. (1982). The Disappearance of Childhood. London, W.H. Allen.

Rubene, Z., Krūmiņa, A., Vanaga, I. (2008) Ievads mediju pedagoģijā. Rīga, RaKa.

Sinhart - Pallin, D. (2001) Medienpädagogik. In: Bernhard, A., Rothermel, L. (Hrsg.) Handbuch Kritische Pädagogik. Weinheim und Basel, Beltz Verlag, S. 383–396.

Strenga, A. (1998) LSDSP un 1934. gada 15. maija apvērsums: demokrātijas likteņi Latvijā. Rīga

Wollersheim, H., W. (2000) Kindheit zwischen Keiserreich und Kinderladen – Etwicklung und Wandel der Kindheit von 1910 bis 1970. In: Larass, P. (Hrsg.) Kindsein kein Kinderspiel. Das Jahrhundert des Kindes (1900-1999). Halle, Verlag der Franckeschen Stiftungen, S. 53-74.

Zelmenis G. (2012) Cenzūra un to reglamentējošā likumdošana Latvijā (1918-1934). Latvijas Vēstures institūta žurnāls. 2012–4 (85)

Rietuma, D. (2021) Kino. Nacionālā Enciklopēdija. Latvijas Nacionālā bibliotēka, Rīga. https://enciklopedija.lv/skirklis/7901-kino

Dimants, A. (2022) - Jaunākās ziņas. Nacionālā Enciklopēdija. Latvijas Nacionālā bibliotēka, Rīga. https://enciklopedija.lv/skirklis/63866

Periodicals

Krolls, O. (1930, September 27th) Kultūrfilma un skola, Kino, Nr. 83. National Library of Latvia, Riga

Krolls, O. (1934, September 14th) Kā top kultūrfilma? Magazina Nr. 122. National Library of Latvia, Riga

Olis, M. (1932, December 7th). Filmu iespaida pētījumi. Mūsu nākotne, Nr. 44. National Library of Latvia, Riga

Paas, H. (1938, May 8th) .1938. Gādāsim kino izrādes skolu jaunatnei. Jaunākās ziņas, Nr. 174. National Library of Latvia, Riga

Paidagogs (1930, September 6th), Kultūrfilma. Filma un Skatuve, Nr. 8. Rīga

P.R. (January 27th, 1932) Kā labāk izmantot kultūrfllmu. Mūsu Nākotne, Nr. 4. National Library of Latvia, Riga

Salnais, Ģ. (1937, July 5th) Ko slēpj mūsu zeme? Atpūta, Nr.653. National Library of Latvia, Riga

Soste, M. (1937, January 30th) Veiksmīgs pusgads mūsu skolās. Brīvā Zeme, Nr. 24. National Library of Latvia, Riga

Štāls M. (1939, March 1st). Filma kā mācību līdzeklis. Audzinātājs, Nr.3. National Library of Latvia, Riga

V.G. (1937, July 13th). Latviešu uzņēmība Parīzes izstādes spogulī. Rīts, Nr. 137. National Library of Latvia, Riga

Vairāk vērības kultūrfilmai (1931, December 16th). Unidentified newspaper clipping (Mūsu nākotne, Nr.33) . National Library of Latvia, Riga

Skolām ieteicams apmeklēt kultūrfilmas (1936, 4th of September). Unidentified newspaper clipping (Brīvā Zeme, Nr.200). National Library of Latvia, Riga

Izstrādā skolu filmu centrāles projektu (1938, October 17th). Unidentified newspaper clipping (Brīvā Zeme, Nr. 236). National Library of Latvia, Riga

Skolas apgādās ar filmu aparātiem (1939, May 26th). Unidentified newspaper clipping (Rīts, Nr. 145). National Library of Latvia, Riga

Kultūras nedēļas skolu izstādes Rīgā (1939, February 1st). Unidentified newspaper clipping (Ministry of Education monthly paper). National Library of Latvia, Riga

Latkino iekšējā un ārējā darbība (1929, November 30th). Unidentified newspaper clipping (Kino, Nr.45). National Library of Latvia, Riga

Archival documents

Celms, J. (1939, June 7th) Correspondence. Ministry of Education, School Department. Department of Foreign Trade of the Ministry of Finance, (Fund nr. 6637, archive nr. 166, A-1200),

Celms, J. (1939, June 14th) Correspondence. Ministry of Education, School Department. State Electrotechnical Factory, Fund nr. 6637, archive nr. 166, A-1238), Latvian State Historical Archive, Riga, Latvia

Dzelme, H. (1940) Ministry of Education, Cinematographer's activity report (1940) (Fund nr. 6637, archive nr. 180), Latvian State Historical Archive, Riga, Latvia

Dzelme, H. (1939) Film technique and methodology courses for teachers. Ministry of Education, School Department. Correspondence with the Latvian Embassy abroad, foreign film studies, etc. on the provision of schools with educational films, cinematic techniques, film lists. (Fund nr. 6637, archive nr. 166), Latvian State Historical Archive, Riga, Latvia

Ozoliņš K., Celms J. (1939, March 8th) Correspondence. Ministry of Education, School Department. Correspondence with the Latvian Embassy abroad, foreign film studies, etc. on the provision of schools with educational films, cinematic techniques, film lists. (Fund nr. 6637, archive nr. 166, A-422), Latvian State Historical Archive, Riga, Latvia

Rosa Veronika Suviranta
University of Helsinki, Finland

Exploring Prompting Strategies for Multimodal Large Language Models in Diagram Classification

Thursday, 06/Mar/2025 3:40pm - 4:00pm
ID: 261 / Session SP 06: 6
Short paper (abstract) | 15-minute presentation with a 5-minute Q&A
Keywords: multimodal large language models, multimodal data, annotation

Multimodal Large Language Models (MLLMs) are gaining interest as potential tools for automating multimodal data annotation in digital humanities, helping to expand annotated corpora, which are essential for enabling large-scale analysis of multimodal communication. MLLMs are an extension of generative Large Language Models, able to process multiple data modalities, such as text, images, and audio. By integrating information across these modalities, MLLMs are able to perform tasks such as image classification, captioning, and visual question answering.

This study explored GPT-4o, a state-of-the-art MLLM, for classifying diagram types, such as cross-sections, cycles and networks and examined how different prompting strategies, including zero-shot, chain-of-thought (CoT), and few-shot prompting, affect its classification performance. In the few-shot condition, the model received a labelled image example from the target category, allowing it to compare new inputs against a visual reference.

The results show that few-shot prompting led to better classification performance than zero-shot and CoT, suggesting that labeled visual examples improve the model's ability to classify inputs. However, performance remained weak across all prompting conditions, highlighting the model's limitations in processing image data.

Firstly, the model seems to lack the ability to correlate high-level concepts with visual patterns. Moreover, the model struggles to distinguish nuanced structural variations between diagrams that have more ambiguous features. Additionally, even with the visual example the model fails to recognise key features that distinguish various diagram types.

These findings suggest that while prompting influences MLLM performance, the improvements remain limited. Although providing more examples might improve performance, model fine-tuning or architectural enhancements may be necessary to make these models useful for multimodal annotation.

Bibliography

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. Advances in neural information processing systems, 33, 1877-1901.

Chen, L., J. Li, X. Dong, P. Zhang, Y. Zang, Z. Chen, H. Duan, J. Wang, Y. Qiao, D. Lin, et al. (2024). Are we on the right way for evaluating large vision-language models? arXiv preprint arXiv:2403.20330 .

Hiippala, T. (2023). Corpus-based insights into multimodality and genre in primary school science diagrams. Visual Communication, 1–22.

Hiippala, T., M. Alikhani, J. Haverinen, T. Kalliokoski, E. Logacheva, S. Orekhova, A. Tuomainen, M. Stone, and J. A. Bateman (2021). AI2D-RST: A multimodal corpus of 1000 primary school science diagrams. Language Resources and Evaluation 55 (3), 661–688.

Kembhavi, A., M. Salvato, E. Kolve, M. J. Seo, H. Hajishirzi, and A. Farhadi (2016). A diagram is worth a dozen images. In Proceedings of the 14th European Conference on Computer Vision (ECCV 2016), Cham, pp. 235–251. Springer.

Kojima, T., Gu, S. S., Reid, M., Matsuo, Y., & Iwasawa, Y. (2022). Large language models are zero-shot reasoners. Advances in neural information processing systems, 35, 22199-22213.

OpenAI (2024). Hello GPT-4o. https://openai.com/index/hello-gpt-4o/. Accessed: 2024-22-10

Rahmanzadehgervi, P., L. Bolton, M. R. Taesiri, and A. T. Nguyen (2024). Vision language models are blind. In Proceedings of the Asian Conference on Computer Vision, pp. 18–34.

Smits, T. and M. Wevers (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.

Taesiri, M. R., T. Feng, C.-P. Bezemer, and A. Nguyen (2024). Glitchbench: Can large multimodal models detect video game glitches? In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 22444–22455.

Tong, S., Z. Liu, Y. Zhai, Y. Ma, Y. LeCun, and S. Xie (2024). Eyes wide shut? exploring the visual shortcomings of multimodal LLMs. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 9568–9578.

Yin, S., C. Fu, S. Zhao, K. Li, X. Sun, T. Xu, and E. Chen (2024). A survey on multimodal large language models. National Science Review , nwae403.

Máté Szentkereszti, **Eszter Mihály, Emese Varga**
Hungarian National Museum Public Collection Centre, National Széchényi Library, Digital Humanities Centre

Visualising, sharing and linking historical and literary knowledge on dHUpla, the platform of Digital Humanities Centre in National Széchényi Library, Budapest

Wednesday, 05/Mar/2025 4:20pm - 4:40pm
ID: 157 / Session SP 03: 2
Short paper (abstract) | 15-minute presentation with a 5-minute Q&A
Keywords: data visualisation, digital philology, National Széchényi Library

In the age of one-minute, stimulus-rich videos, it is a major challenge to capture people's attention with scientific topics. Scholars working in humanities are often unable to present historical, literary and linguistic research based on sources held in public collections in an engaging way. The ars poetica of digital philologists at the National Széchényi Library in Budapest claims that, paradoxically, it is the digital humanities branch of philology, whose analogue counterpart is an isolated discipline within the humanities, that can help reach non-academic audiences and connect them with researchers working with fundamentally analogue methods. Our paper claims, that one of the best ways to coming out of the ivory tower is possible through data visualisations. For this very reason, the focus will be on the role of visualisations in our work on dHUpla, the platform of Digital Humanities Centre (DHC), as well as the tools that bring these visualisations to life.

Since the establishment of the DHC in 2020, our fundamental aim has been to raise awareness among researchers and the general public about the manuscript material in public collections, and the unique items in the collection of the National Széchényi Library (NSZL). The used corpus usually consists of writers' and intellectuals' correspondence, and other historic sources, which will be expended with more sources (for example egodocuments, just like diaries, lecture notes and testaments). The paper will introduce our workflow (Collecting resources, Structuring data, Setting parameters, Export and Display, Finalizing data structure, Publishing) which may provide guidance to GLAM sector institutions to create graph-based, online, and freely accessible visualisations of large data sets.

Through their dual function, data visualisations can be an effective tool for the promotion of science and culture. Data stored in databases (catalogues, library databases, tables, knowledge repositories) can raise new questions and new perspectives in research through an impactful visual representation. Based on the correspondence, data visualisations reconstruct the relationship system of a person in focus, revealing previously unknown characteristics and aspects. So far, we have chosen important, well known and/or more forgotten figures of Hungarian culture and history such as writer Zsigmond Móricz (1879-1942), poet Sándor Petőfi (1823-1849), literary journal editor József Kiss (1843-1921), and the founder and patron of our library, Ferenc Széchényi (1754-1820). The finished data visualisations of their correspondence show, for example, how many people wrote letters to Ferenc Széchényi in Hungarian at the beginning of the 19th century, or how many letters were written by men and women to József Kiss, the famous editor, and how many of these were professional letters. The paper give insight, what kind of information and how can be derived from visualisations and how can be used in academic research and in public education as well. The interactive, searchable graphs, family charts and maps link our texts published on dHUpla together, and can be used as tools in games such as our virtual escape room 'The journey of the manuscript', which will be presented in this paper. Finally, our brand-new initiative to reconstruct Mór Jókai (1825-1904), one of the most famous Hungarian writer's garden with AI (based on objects, manuscripts, graphics) will be introduced, showcasing NSZL DHC's multidimensional program in knowledge dissemination.

Bibliography

dHUpla website: https://dhupla.hu/

Mikhail Tamm[1], Mila Oiva[2], Ksenia Mukhina[1], Mark Mets[1], Maximilian Schich[1]
[1]TLU, Estonia; [2]University of Turku, Finland

Quantifying world geography as seen through the lens of Soviet propaganda

Thursday, 06/Mar/2025 10:00am - 10:30am
ID: 213 / Session LP 04: 3
Long paper (abstract) | 20-minute presentation with a 10-minute Q&A
Keywords: historical news corpus, quantitative geography, statistical analysis, urban scaling, history of the Soviet Union

Introduction

Large scale cultural data is subject to quantitative patterns, which in many cases lend themselves to measurement and interpretation. Representation and selection biases are among the most important factors producing those patterns. One particular example is unequal representation of geographical locations in the media. The ability to pinpoint and quantify it can give essential insights into the underlying normative worldview of the media-producing society.

While geographical and spatial biases are present at all spatial scales, from continents to neighbourhoods, cities form a natural probe to study representation of geographical space in historical media. Cities are numerous, their size is relatively well-defined, spans multiple orders of magnitude, and is reasonably well-documented historically. Recent influx of ideas from complex systems theory into urban science, especially the idea of urban scaling [1-5] provides a useful conceptual framework for understanding the city representation.

We develop a general procedure for extracting insights regarding the representation of geographical space in historical media from the data on how cities are mentioned in a historical news corpus. Our method consists of following feedback-loop-forming steps: (i) formulation of a hypothesis about parameters governing city representation; (ii) calculation of the parameters of the hypothesis by minimization of an explicitly defined loss function, (iii) elimination of irrelevant parameters based on a predetermined information-theoretic criterion, and (iv) correction of the hypothesis based on qualitative analysis of the outliers.

We exemplify this procedure by the systematic study of the corpus of Soviet newsreels ``Novosti Dnya'' (News of the day)[6]. Newsreels -- short news films shown in cinemas before the feature film -- were influential means of depicting the world in the 20th century. Throughout almost all history of the Soviet Union, the newsreels were heavily censored. Their content thus reflects the prescribed worldview, the set of topics, places and individuals, which were considered appropriate to be presented and discussed in an official source. They provide therefore an interesting glimpse into the history of the Soviet Union and its political and media culture.

Despite the official Soviet ideology of equality, interconnected social and spatial hierarchies were at the core of the Soviet system. These hierarchies originated in both the political ideology and the pragmatic considerations of usefulness for the state and where entangled with spatial hierarchy with Moscow at the very top[7-9]. Meanwhile, representation of the outside world was determined by current politics, and its shifting tendencies of isolationism or expansionism [10-11].

Methods

The method we develop here aims to extract the quantitative estimates of the factors determining the frequency of mentions of the cities in a robust and reliable way. Input consist of a list of N cities with the numbers of times n_i ($i = 1,..., N$) they are mentioned, and a *hypothesis*, i.e. a rule defining the *expected number* of mentions of each city m_i ($i = 1,..., N$) . The observed number of mentions is then interpreted as a random number with average m_i [12], allowing to construct a *loss function* -- a mathematical expression of how unlikely is to see the observed number given the expected one. By minimizing this loss function we can optimize the parameters of the hypothesis. Moreover, we add two procedures to systematically improve the hypotheses: one avoids overfitting by removing irrelevant parameters (see [13]), another allows, by qualitative study of outliers, i.e. cities for which expectation is most diverging from observation, to include additional aspects, which are initially overseen.

Data

We use the corpus of the Soviet Newsreel ``Novosti Dnya'' (News of the Day) [6], consisting of over 1700 short films, split into 12,707 stories with short text outlines [14]. The corpus is almost complete for 1954--92 with some additional issues from 1944-53. Cities are included in the list of cities of interest if they exceed a preset population threshold [14]. The mentions of each city where manually classified by fluent Russian-speakers into 5 categories: (i) direct mention of a city and city-dwellers, (ii) mention of organizations and industrial enterprises located in the city and named after it, (iii) mentions of the region surrounding the city, and organizations located there, (iv) entities named after the city but located elsewhere, (v) homonyms and coincidences. Only mentions of type (i) and (ii) are included in the analysis.

Results

We construct a series of models explaining representation of the cities separately inside and outside the USSR. We start with a hypothesis that city representation is determined by its population, and repeatedly refine this hypothesis by studying the outliers, while simultaneously pruning out irrelevant variables to avoid overfitting.

We find that cities inside USSR are mentioned much more (by a factor of several hundreds) than similar-size cities outside, in both cases representation on average grows superlinearly with city size: doubling the size of a city leads to approximately 2.3 times growth in representation.

For Soviet cities there are 7 city specializations which statistically significantly change the city representation, 5 of them -- being a capital of Union-level republic, being a port on the Black, Baltic sea or the Pacific coast, having a huge (>2 MW) hydroelectric dam, a full-cycle steelwork or a non-ferrous metallurgy plant -- boost the number of mentions by 55% - 110%, while 2 others -- being a capital of autonomous national republic inside Russia proper and specializing in coal-mining -- suppress the number of mentions by 25-40%.

Moreover, there are geographical parts of the Soviet Union, which are systematically over- and under-represented, the former being Northern Kazakhstan and the close vicinity of Moscow, the latter being Western Siberia, Volga Region, Eastern and Central Ukraine, but most prominently Donbas, Western Urals and Asian republics of the USSR (except for Georgia and Northern Kazakhstan).

The dataset for foreign cities is more sparse. However, we were able to identify two main factors: the capital status of a city and its geographical location. Interestingly, boost for the capital status depends on the population of the corresponding country: capitals of larger countries get a larger boost. In terms of geographical location, we expected to see different levels of representation for socialist, developed capitalist and developing countries. However, the optimization resulted in a more fine-

grained picture: while optimization algorithm indeed grouped developing countries into a single bucket, it split socialist and capitalist camps into four groups each.
That is, cities of the most loyal socialist countries (Bulgaria, Czechoslovakia and Mongolia) are mentioned roughly 60 times more than equivalent-size cities in the developing world, those in Poland, East Germany and Hungary -- 35 times more, while cities from non-European socialist countries (except Mongolia) and from ``problematic'' Yugoslavia and Romania -- only 15 times more. In turn, cities from USA, Canada and Australia are mentioned 2.5 times more than developing world equivalents, those in most of capitalist Europe - 4.5 times more, and those in neutral Finland and Austria - amazing 75 times more. Finally, socialist China and capitalist Japan show now over-representation at all compared to the baseline of developing countries.

Discussion

Full interpretation of these biases needs further qualitative analyses of the corpus, coupled with other topical historical sources. However, we observe the following important repeating motives.

Our corpus shows a clear hierarchy of representation with the Soviet Union on top, followed by the Socialist block, the developed capitalist world and, finally, the developing world. Representation of cities grows superlinearly with city population, indicating positive agglomeration effects, and is boosted by a capital status.

Contrary to the messaging of the official Soviet ideology, which emphasized equality of nations and anticolonial movement, the silently sold Soviet worldview is heavily centered on Europe being in the role of a privileged or hierarchically higher "Other"[15]. In agreement with previous qualitative observations [7,9,16,17], we find that European countries (both socialist and capitalist) are mentioned more than their counterparts elsewhere, while western regions of the USSR are mentioned much more than Central Asia and Southern Caucasus.

Some particular regions and branches of heavy industry have an outsized ideological importance. Regional examples are Northern Kazakhstan inside the USSR, the most loyal countries of the socialist block, and, most strikingly, the two neutral capitalist countries, Austria and Finland.

Seemingly, Soviet worldview deliberately avoids mixed and intermediate cases and situations: while a trait is celebrated and emphasized in its fully developed form (superlarge dams, Far East location, Union-level capital status), intermediate forms of the same trait (medium-sized dams, location in West Siberia or Urals, capital of a lower-level national autonomy) are often under-represented. Possibly a similar mechanism is behind the under-representation of Eastern and Central Ukraine: while Eastern Europe, including republics of the Soviet Western frontier, is overwhelmingly important for the Soviet worldview, Eastern and Central Ukraine with its mixed Ukrainian-Russian population might seem neither fully Eastern European nor fully Russian. If true, the interplay of these two factors: over-fixation on Eastern Europe and denial of the fact that Ukraine fully belongs to it, might be instructive in understanding the worldview which led to the current Russian aggression against Ukraine.

While studying a particular example of a Soviet media corpus, we develop a general approach to extracting information on geographical biases from historical news corpora. The suggested procedure combines quantitative and qualitative steps into a single feedback loop, allowing to systematically refine hypotheses about relevant factors and to measure biases in robust quantitative way. The methodology developed here can be used for the analysis of multiple other datasets and hopefully will become a standard in the field.

Bibliography

[1] D. Poumain, Scaling laws and urban systems, Working papers of the Santa Fe Institute, 2004.

[2] L.M.A. Bettencourt, J. Lobo, D. Helbing, C. Kühnert, G.B. West, Growth, innovation, scaling, and the pace of life in cities, PNAS, 104, 7301-7306 (2007).

[3] L.M.A, Bettencourt, J. Lobo, D. Strumsky, G.B. West, Urban Scaling and Its Deviations: Revealing the Structure of Wealth, Innovation and Crime across Cities. PLoS ONE, 5, e13541 (2010).

[4] L.M.A. Bettencourt, The Origins of Scaling in Cities. Science, 340, 1438 (2013).

[5] E. Arcaute, E. Hatna, P. Ferguson, H. Youn, A. Johansson, and M. Batty, Constructing cities, deconstructing scaling laws. J. R. Soc. Interface,12, 20140745 (2015).

[6] M. Oiva, K. Mukhina, V. Zemaityte, A. Karjus, M. Tamm et al., A framework for the analysis of historical newsreels. Hum. Soc. Sci. Comm., 11, 1-15 (2024).

[7] N. Pianciola, Stalinist Spatial Hierarchies: Placing the Kazakhs and Kyrgyz in Soviet Economic Regionalization. Central Asian Survey, 36, 73-92 (2017).

[8] E. Dobrenko. Late Stalinism: the Aesthetics of Politics. Yale University Press, 2020.

[9] E.T. Megowan, ‘Writers Live Only in Moscow and Leningrad’? Navigating Soviet Spatial and Cultural Hierarchies, 1941–45. Kritika: Explorations in Russian and Eurasian History, 22, 285-311 (2021).

[10] B. McNair. Glasnost, Perestroika and the Soviet Media. Routledge, 2006.

[11] K.A. Bogdanov, The USSR Instead/inside of Europe: Soviet Political Geography in the 1930s–1950s. Studies in East European Thought, 62, 401-412 (2010).

[12] J.C. Leitao, J.M. Miotto, M. Gerlach, E.G. Altmann, Is this scaling nonlinear? R. Soc. Open Sci., 3, 150649 (2016).

[13] K.P. Burnham, D.R. Anderson, K.P. Huyvaert, AIC model selection and multimodel inference in behavioral ecology: some background, observations, and comparisons, Behav. Ecol. Sociobiol., 65, 23–35 (2011).

[14] M.V. Tamm, M. Oiva, K.D. Mukhina, M. Mets, M. Schich, Supplementary materials for this paper, https://github.com/thummm/soviet_newsreels/.

[15] L.E. Cahoone. Introduction, In: "From Modernism to Postmodernism: An Anthology Expanded". Wiley-Blackwell (2003).

[16] P.Koivunen. Friends, ‘Potential Friends,’ and Enemies: Reimagining Soviet Relations to the First, Second, and Third Worlds at the Moscow 1957 Youth Festival.' In "Socialist Internationalism in the Cold War", P. Babiracki, A. Jersild, eds. (Springer International Publishing, 2016) p. 219–247.

[17] J. Gronow and S. Zhuravlev. A Window to the West and Door to the East. The Tallinn Fashion House as a Part of the Soviet Fashion Design System. In "Mood Ja Külm Sõda - Fashion and the Cold War", (Estonia: Art Museum of Estonia - Kumu Art Museum, 2012) p. 108–137.

Christina Tånnander[1,2], **Jens Edlund**[1]
[1]KTH Royal Institute of Technology; [2]MTM, Swedish Agency for Accessible Media

Demonstration of Sardin, a Swedish Speech-Oriented Text Processing System

ID: 250 / **Poster Session 2: 25**
Poster and demo (full-text) with accompanying a 1-minute lightning talk
Keywords: speech-oriented text processing, speech technology, text-to-speech synthesis

We present Sardin, a Swedish speech-oriented text processing system designed primarily to prepare text for text-to-speech (TTS) synthesis. Sardin processes input documents in XML, EPUB, or TXT formats through multiple modules, ultimately producing an SSML string suitable for use with TTS systems and similar applications. This paper highlights potential applications for Sardin and serves as a demonstration, providing brief descriptions of the system's modules alongside illustrative examples. The modular nature of Sardin makes it suitable for a range of research and development tasks outside of pure TTS synthesis, which is also exemplified in the demonstration.

Melissa Terras[1,2], Bettina Anzinger[1], Günter Mühlberger[1,3], C. Annemieke Romein[1,4,5], Andy Stauder[1], Florian Stauder[1]
[1]READ-COOP SCE, Austria; [2]University of Edinburgh; [3]University of Innsbruck; [4]University of Twente; [5]University of Bern

From Public Funding to Platform Cooperative: The READ-COOP Model for Sustainable Digital Scholarship

ID: 315 / Poster Session 2: 33
Late-breaking poster
Keywords: AI/ML, Handwritten Text Recognition, Digital Infrastructure, Cooperative Models, Democratize Technology

How can we sustainably build digital scholarship infrastructures that serve their communities while encouraging co-ownership and development input? This question is addressed through examining READ-COOP, the first platform cooperative to develop and host its own AI and Machine Learning tools. READ-COOP is a European Cooperative Society hosting Transkribus, an Automated Text Recognition platform that unlocks historical documents with AI, winning the European Union's Horizon Impact Award 2020. By November 2024, it had 229 members from 35 countries, including leading libraries, archives, and universities, co-owning an infrastructure that has generated accurate transcriptions of over 100 million digital images of historical texts, with more than 300,000 registered users.

READ-COOP and Transkribus emerged from two European Commission funded projects. TranScriptorium (2013-15, €2.4m) produced the Machine Learning pipeline for generating accurate transcriptions from digital images of historical texts. The Transkribus Graphical User Interface launched in 2015 as a downloadable Java-based client programme. The Recognition and Enrichment of Archival Documents (READ) project (2016-2019, €8.2m) further developed the service as a technology hub. By 2019, after €10.6m in funding, the platform achieved Character Error Rates below 5% for handwritten text and 1% for print material.

In the current landscape, social, political, and economic dimensions of data are controlled by few monopolies focused on capital accumulation and extraction. However, READ-COOP demonstrates that cooperative infrastructures can provide an alternative to extractive shareholder-oriented capitalism. The cooperative model supports democratic decision-making while enabling revenue generation for infrastructure and service improvements.

The cooperative approach is particularly suitable for digital infrastructures initially developed through public funding, provided they have a substantial, defined, and engaged user base. This model represents a significant shift in how we conceive, implement, and sustain systems, advocating for a more democratized technology landscape. Cooperatives, though underutilized as a legal framework and business model, show potential for funding specific digital infrastructures and supporting the creation of trustworthy, responsible AI in various settings beyond the creative and cultural heritage sector.

READ-COOP's success provides a blueprint for establishing other cooperative digital infrastructures and suggests an alternative future for responsible AI cooperatives. This approach could be particularly valuable for digital scholarship, digital cultural heritage, digital humanities, and AI tools, products, services, and platforms. The cooperative business model offers a promising alternative to traditional funding structures, especially for projects transitioning from public funding to sustainable operation. Through READ-COOP's example, we see the potential for cooperative models to create and maintain digital infrastructures that truly serve their communities while ensuring long-term sustainability and ethical governance.

Bibliography

Cheffins, B. R. (2021). “Stop Blaming Milton Friedman!”. Washington University Law Review, Volume 98, Number 6, 2021. HeinOnline, https://heinonline.org/HOL/P?h=hein.journals/walq98&i=1629.

EU Science & Innovation (2020). “TRANSKRIBUS, winner of the Horizon Impact Award 2020”. YouTube, September 23rd, 2020. https://www.youtube.com/watch?v=CL1fe3wwOaI

Huberman, J. (2022). The spirit of digital capitalism. John Wiley & Sons.

Muehlberger, G., Seaward, L., Terras, M., Ares Oliveira, S., Bosch, V., Bryan, M., Colutto, S., Déjean, H., Diem, M., Fiel, S., Gatos, B., Greinoecker, A., Grün- ing, T., Hackl, G., Haukkovaara, V., Heyer, G., Hirvonen, L., Hodel, T., Jokinen, M., Kahle, P., Kallio, M., Kaplan, F., Kleber, F., Labahn, R., Lang, E. M., Laube, S., Leifert, G., Louloudis, G., McNicholl, R., Meunier, J. L., Michael, J., Mühlbauer, E., Philipp, N., Pratikakis, I., Puigcerver Pérez, J., Putz, H., Retsinas, G., Romero, V., Sablatnig, R., Sánchez, J. A., Schofield, P., Sfikas, G., Sieber, C., Stamatopoulos, N., Strauß, T., Terbul, T., Toselli, A. H., Ulreich, B., Villegas, M., Vidal, E., Walcher, J., Weidemann, M., Wurster, H., Zagoris, K. (2019). Transforming scholarship in the archives through handwritten text recognition: Transkribus as a case study. Journal of Documentation, 75(5):954–976, January 2019. ISSN 0022-0418. doi: 10.1108/JD-07-2018-0114. URL https://doi.org/10.1108/JD-07-2018-0114.

Sánchez, J. A., Bosch, V., Romero, V., Depuydt, K., and De Does, J. (2014). Handwritten text recognition for historical documents in the transcriptorium project. In Proceedings of the first international conference on digital access to textual cultural heritage (DATeCH ‘14) (pp. 111-117). https://doi.org/10.1145/2595188.2595193

Schneider, N. (2021). Enabling Community-Owned Platforms. A Proposal for a Tech New Deal. In Moore, M., and Tambini, D. (2021). Regulating Big Tech: Policy Responses to Digital Dominance. Oxford: Oxford University Press.

Scholz, R. T. (2023). Own This!: How platform cooperatives help workers build a democratic internet. London: Verso Books.
Schon, D.A. (1983) The Reflective Practitioner: How Professionals Think in Action. New York, NY: Basic Books.

Terras, M., Anzinger, B., Gooding, P., Muehlberger, G., Nockels, J., Romein, C. A., Stauder, A., Stauder, F., (Forthcoming 2025). The artificial intelligence cooperative: READ-COOP, Transkribus, and the benefits of shared community infrastructure for automated text recognition. Accepted, Open Research Europe.

Université de Montreal (2018). Montreal Declaration for a responsible development of artificial intelligence. (2017, November, 3). Announced at the conclusion of the Forum on the Socially Responsible Development of AI. https://www.montrealdeclaration-responsibleai.com/

Darja Tokranova
Tallinn University, Estonia

Social Media as Redefined Scene for Estonian Visual Arts

ID: 111 / S01: 1
Doctoral Consortium
Keywords: Estonian Creative Economy, Social Media, Non-Fungible Tokens, Art.

Social Media challenges the rigid institutional format of the art market, where beginners or amateurs are traditionally deprived of certain privileges of the artistic community and must compete for their right to belong in the industry by jumping through certain hoops (exhibitions, grants, tiers of official education, peer recommendations etc.). Social networks can make executing, exhibiting and selling art easy: create and upload, show the world and make a sale: with no curators, museums, or institutions involved. To study the issue further and compile it in the form of an article-based PhD dissertation author aims to explore three different perspectives of creatives and their relationship with Social Media: practices of Social Media use among the Nordic visual and audiovisual creatives, author's own art-project- based experience; and how Estonian practitioners experience Social Media within their work.

This In the year 2006, P. Weibel expressed a view that new media has impacted the entirely new art situation, including traditional arts: "No-one can escape from the media." In the year 2023 this can be rephrased as "No-one can escape from the social media", as we enter a new era where having a digital artistic identity becomes natural for the creative practitioner.

Visual and audiovisual materials (including, but not limited to graphics, sound compositions, videos and clips, recordings of transmitted installations and performances) are currently the most shared elements in Social Media (SM). [1] The rapid expansion of the latter brought a paradigm shift in the domain of audiovisual arts, influencing creation, exhibition and curation of art. Roles that in the traditional art world, were reserved for those with advanced degrees in art history, are now accessible to members who have no prior knowledge or education in arts. [2] Previously a "niche" discipline, based on the Institutional Theory of Art introduced by A. Danto [12] and developed by G. Dickie [13] where museums acted among the main venues of the 'gate-keeper' system that distinguished between high and low art [3], isolated from mainstream and accessible only by a limited circle of insiders and enthusiasts, it is now being exposed to the world and transformed by accelerating technological progress, where the secondary activity of networking is absolutely vital. [4] Year 2014 was a turn-around point for Social Media being used as an art platform and art medium at the same time. The installation "New Portraits" by Richard Prince consisted of enlarged photos from the artist's Instagram feed and created opinion-rich resonance when photos started selling for as much as 100,000$ at Frieze Art Fair New York. [5]

In the same year Amalia Ulman created a performance "Excellences & Perfections ", that lasted several months and gathered an audience of 90000 followers. She has created what critics heralded as the "first Instagram masterpiece" and in 2016 the piece was included in a group show at Tate Modern, "Performing for the Camera", making her the first Social Media artist to enter a top institution. [6] Already, a couple of years later, in 2016, over 80 percent of all Generation Y art buyers bought fine art online, with almost half of online buyers using Instagram for art-related purposes [8]. Looking at the worldwide examples one may wonder about the state-of-the-art in Nordics and Estonia in particular.

This study uses Nordics as a contextual background including the following countries: Iceland, Norway, Denmark, Sweden, Finland and then focuses on Estonia in particular. Estonia is a country with a small density population but is one of the leading countries in Europe in digital literacy and Internet penetration among the population [9], hence the use of Social Media within the population is quite advanced - in the tear of 2017 around 72% of Estonian Internet users were also using Social Media [11]. The Creative Economy of Estonia has a considerable socioeconomic dimension, according to Josing et al. (2022) around 28,300 employees worked in the creative industries sector in 2019 in Estonia, which made up 4.2% of employed persons in Estonia. One might speculate that social media may positively contribute to artistic publicity, increased attention to one's works, better sales and export of talent. It also has the potential to lessen the pressure to interact with traditional art market actors such as: various artistic associations and clubs, formal education institutions, curators, art critic communities, museums and exhibition spaces, distributors of grants and hence, provides more artistic freedom and less pressure to chase the traditional artistic career paths.

This research is aimed to gain a better understanding of state-of-the-art and explore how Social Media affects Estonian Creative Economy (with focus on Audiovisual sub-sector) and how local creatives use social networks (Instagram, Facebook, TikTok, Twitch etc.) to create, exhibit and profit from their art. Following research questions are being explored:

RQ1: How is Social Media used among visual and audiovisual creatives in Nordics?

RQ2: How is the process of creating and promoting an art project on social media reflected in personal experience through an autoethnographic lens?

RQ3: How is Social Media used among visual and audiovisual creatives in Estonia?

This research has the potential to serve the interest of multiple stakeholders. Some of the Estonian academics point to the need for more scholar interest towards local creative industry [14] and the same sentiment is shared on a governmental level within the framework of "Kultuur 2030" development programme [15]. It is also on par with the European Commission's flagship programme to support the culture and audiovisual sectors "Creative Europe" [16], where both academic research activities and audiovisual sector are given special attention.

Bibliography

Darja Tokranova (2024). Contributing to the Environment with Tackling UX Issues: European Environmental Agency Case-Study. 1−3. Trinidad Wiseman Consultancy Blog, 2024.

Darja Tokranova (2024). When Four Becomes One: Things to Consider When Merging Complex Systems Together, ECHA Case Study. 1−4. Trinidad Wiseman Consultancy Blog, 2024.

Bauters, Merja Lina, Darja Tokranova, Liyanachchi Mahesha Harshani De Silva, Juri Mets. (2023). The Exploration of Skill Gaps and Ecosystem Potential among Estonian Creatives. Sustainability, 15 (18), 13687. DOI: 10.3390/su151813687.

Tokranova, Darja; Bauters, Merja Lina (2022). Applying participatory design principles to collaborative art-creating sessions. 25th International Academic Mindtrek conference. ACM, 347−353. DOI: 10.1145/3569219.3569367.

Tokranova, Darja (2020). Tackling Ethical Implications of Mobile Banking Product Development Through the Value Sensitive Design Approach. NordiCHI '20: Proceedings of the 11th Nordic Conference on Human-Computer Interaction: Shaping Experiences, Shaping Society: Tallinn, Estonia, October 25 - 29, 2020. ACM , Article No 113. DOI: 10.1145/3419249.3420072.

Mõttus, Mati; Lamas, David; Tokranova, Darja (2014). Evaluating Aesthetics in Episodical Interactions. In: MIDI '14 Proceedings of the 2014 Mulitmedia, Interaction, Design and Innovation International Conference on Multimedia, Interaction, Design and Innovation. (1−14). New York, NY, USA: ACM. DOI: 10.1145/2643572.2643583.

Darja Tokranova (2014). The Effects of User Interface Aesthetics in the User's Experience. (Magistritöö, Tallinna Ülikool). Tallinn: Tallinna Ülikool.

Darja Tokranova, Merja Lina Bauters
Tallinn University, Estonia

Examining Social Media Utilization Among Estonian Creatives

Friday, 07/Mar/2025 1:30pm - 1:50pm
ID: 110 / Session LP/SP 09: 1
Long paper (abstract) | 20-minute presentation with a 10-minute Q&A
Keywords: social media, influencers, art, Estonian creative economy

Social media poses a challenge to the entrenched institutional framework of the art market, which historically was a “niche” discipline, with museums among the main venues of the 'gate-keeper' system that distinguished between high and low art (Varnedoe et al., 1990). The advent of social media presents a paradigm shift by affording individuals the capability to create, exhibit, and sell artwork by simply disseminating content online, thus circumventing traditional gatekeepers such as curators, museums, and institutions. At the same time, contributors are also challenged with the stigma of use of social media - narcissism being among the motives (Nabi, 2017).

The creative industries constitute a significant component of the Estonian economy, employing approximately 4.2% of the population in 2019 (Josing et al., 2022), with the industry players increasingly integrating digital technologies into artistic practices (Bauters et al., 2023). The ways local artists communicate with spectators of their art are also gradually expanding to the digital space: many established contributors to the Estonian creative economy, such as pop-musicians Nublu and Tommy Cash or visual artists Mark Kostabi and “Estonian Banksy” graffiti artist Edward von Lõngus, now have either dedicated web pages or social media accounts in an attempt to stay closer to their audience. Estonian artists are gaining an increasing presence and leverage in Estonia and abroad, promoting political causes and social issues, and thereby become political influencers (Riedl et al, 2023). An urban art performance, “Hetk” (“Moment” in Estonian), launched in February 2024 by Estonian artists Estookin Andreen and Rebeca Parbus, sparked numerous heated discussions across social media platforms such as X (ex-Twitter), Facebook and others. They presented manipulated images of Tallinn city urbanistic locations using Adobe Photoshop, portraying post-war buildings engulfed in smoke and damage (Soans et al., 2024) next to the actual locations. This "before and after" comparison juxtaposed real-life scenery with a somber depiction of a potential wartime scenario. Authors encountered both critiques and support for their bold ways of artistic expression (Aadli et al., 2024), illustrating the potential of Estonian artistic influence to resonate with both domestic and international audiences.

Darja Tokranova, Merja Lina Bauters
Tallinn University, Estonia

Reflecting on Digital Art Value: NFTs' Potential for the Art-Market Parity

Wednesday, 05/Mar/2025 1:30pm - 2:00pm
ID: 140 / Session LP 03: 1
Long paper (full-text) | 20-minute presentation with a 10-minute Q&A
Keywords: Blockchain, Non-Fungible Tokens, Digital Art, Creativity

Until recently, digital art was perceived as something of secondary value compared to the physical artistic artefacts. One of the reasons for that being its predisposition for duplication and hence the inability to assign “original artwork” label to the digital file and represent it as a unique object on artistic market. But the rapid pace of popularization of blockchain technologies in creative communities through the use of Non-Fungible Tokens has a seeming potential to change the perception of digital art. The ERC721 standard sets a precedent for authentication and traceability of digital artworks suggesting that the old paradigm might shift, and digital art will gain value and attention comparable with traditional fine art. In this article we discuss the problematics of digital art representation on art market and the issue of digital creations' pricing. We use photo stocks and print-on-demand platforms as an example for pre-NFTs digital art monetization. We then discuss the changes caused by Non-Fungible Token blockchain technology in the digital art market in recent years and the implications that come with the change. We then illustrate theoretical tenets with expert interview that suggest that while successful NFT projects offers publicity and profit to the creators, the level of complexity and unpredictability of results sets a high bar for entering the market.

Tanel Torn
Estonian Literary Museum

Classification of Instrumental Folk Music Recordings with Machine Learning

ID: 292 / **WS10: 7**
Tradition Archives Meet Digital Humanities II
Keywords: folk music, archive recordings, machine learning, audio analysis, digital ethnomusicology

The analysis of traditional folk music often involves labor-intensive manual processes, such as transcription by hand, limiting the ability to explore large archival collections. As digital archives of folk recordings continue to grow, there is a pressing need for automated approaches to organize, classify, and understand these culturally significant materials. This study explores the potential of machine learning techniques to address these challenges, with a focus on practical methods for analyzing audio recordings directly without relying on transcription.

Machine learning models operate by identifying underlying structures and relationships within audio features, enabling the automatic classification of folk music recordings into distinct groups based on their melodic, harmonic, and rhythmic properties. While these models work with complex, latent patterns that may not be directly interpretable by humans, they facilitate efficient handling of vast datasets and help uncover trends that would otherwise be challenging to detect. This approach also aids in identifying inconsistencies within digital archives, such as mislabeled or duplicate recordings.

The Estonian Folklore Archive, with its rich collection of nearly 10,000 recordings from the 20th century, spanning multiple regions and time periods, provides a valuable testbed for this research. The recordings, captured using diverse equipment from early phonographs to modern digital recorders, exhibit significant variability in quality and timbre, reflecting the technological and cultural diversity of the archive.

This presentation outlines preliminary findings and ongoing work to evaluate machine learning's applicability to ethnomusicological research. By automating the clustering of audio recordings, this approach seeks to complement human expertise, enabling researchers to explore cultural heritage more comprehensively and at scale. While results are still emerging, this work aims to demonstrate the potential of these methods in advancing the study of traditional folk music.

Jon Carlstedt Tønnessen, **Magnus Breder Birkenes**
National Library of Norway, Norway

WebData: Research Infrastructure for the Norwegian Web Archive

ID: 126 / Poster Session 2: 12
Poster and demo (abstract) with accompanying a 1-minute lightning talk
Keywords: infrastructure, web archive, data analysis, collections as data, accessibility

Web archives contain petabytes of data with great potential for utilisation in research and knowledge-production. However, scholars have described significant problems to work with web archives, due to a range of technical and legal issues. Currently, the lack of dedicated research infrastructures and services to work with web archives at scale leaves many of these huge data repositories inaccessible to researchers.[1]

In 2025, we will launch WebData: a project to build a state-of-the-art research infrastructure for the Norwegian Web Archive (NWA). Funded by the Research Council of Norway, WebData will provide search in full-text and metadata, access to collections as data, and quality resources for computational analysis of the web archive collection. The aim is to build a user-friendly platform for scholars, students and others in need of finding, discovering, retrieving and analysing data from NWA, adhering to FAIR principles.[2]

The poster will present the main activities of the WebData project for 2025-2029:
- Building a data platform for searching, exploring and retrieving data with search in full-text and granular metadata,
- Ensuring user-oriented development by assessing scholarly problems and needs,
- Data enrichment with high-quality annotation, facilitating NLP-based analysis like named-entity recognition (NER), event detection and sentiment analysis,
- Building a pipeline to pseudonymise sensitive personal data, providing access to data that would otherwise be restricted
- Improving detection of Indigenous languages (Sámi and Kven) and representation of Indigenous content in the Norwegian Web Archive
- Trainings with scholars and students, in cooperation with universities

The poster will include an overall timeline for the WebData project for the period 2025-2029. It will also present how scholars can contribute to its development, and depict how scholars can expect to benefit from the infrastructure when the WebData research infrastructure is production-ready.

The WebData project is led by the National Library of Norway (NB), with the University of Oslo (UiO), The Arctic University of Norway (UiT) and the Norwegian Computing Center (NR) as partners.

Endnotes

[1]: Maemura (2023): 1–14; Tønnessen (2024); Brügger (2021): 217–24.
[2]: 'WebData: Research infrastructure' (2023); Research Council of Norway (2024).

Bibliography

Brügger, Niels. 2021. "The Need for Research Infrastructures for the Study of Web Archives." In The Past Web: Exploring Web Archives, eds. Daniel Gomes, Elena Demidova, Jane Winters, and Thomas Risse, 217–224. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-63291-5_17.

Maemura, Emily. 2023. "All WARC and No Playback: The Materialities of Data-Centered Web Archives Research." Big Data & Society 10 (1): 1–14. https://doi.org/10.1177/20539517231163172.

Research Council of Norway. 2024. "1,3 milliardar kroner til forskingsinfrastruktur."
https://www.forskningsradet.no/nyheter/2024/13-milliardar-forskingsinfrastruktur/ (accessed September 26, 2024).

Tønnessen, Jon Carlstedt. 2024. "Diving into the Digital Heritage. (re)Searching the Norwegian Web Archive." Paper presented at DHNB2024, Digital Humanities in the Nordic and Baltic Countries, [Reykjavik].

Tønnessen, Jon Carlstedt. 2024. "Navigating the Scales of Web Archives: Leveraging Data for Research." [Manuscript submitted for publication].

"WebData: Research Infrastructure for Web Data." [Application to the Research Council of Norway]. 2023.

Iulianna van der Lek[1], **Giulia Pedonese**[2], **Francesca Frontini**[1]
[1]CLARIN ERIC, Netherlands, The; [2]CNR-ILC, Italy

Building a Sustainable CLARIN Trainers' Network: Challenges and Lessons Learnt from the CLARIN-IT Consortium

ID: 294 / **WS09: 4**
Workshop on Digital Humanities and Social Sciences/Cultural Heritage (DHSS/DHCH) in Higher Education
Keywords: research infrastructures, FAIR learning resources, training infrastructure, trainers' network

Aims & Objectives

The main aim of this presentation is to share how CLARIN ERIC, a European Research Infrastructure dedicated to making digital language resources and technologies accessible to SSH research communities, educators and cultural institutions, managed to gradually develop a network for educators, trainers and researchers to support each other and share resources in DHSS/DHCH training events, summer schools and workshops. After briefly introducing the CLARIN trainers' network and discussing the challenges encountered to make this network sustainable, the presentation will focus on the experience of the CLARIN Italian node with developing a training infrastructure platform for their community to support modern teaching practices while managing training materials according to the FAIR principles: Findable, Accessible, Interoperable, and Reusable. More specifically, the presentation will share how we applied the FAIR-by-Design methodology proposed by the Skills4EOSC project to reuse and adapt existing learning content from the CLARIN Learning Hub for other teaching contexts and target audiences.

The modular aggregation of training materials, supported by the FAIR-by-design methodology, allowed us to adapt single modules to specific academic courses. For example, the CLARIN-IT consortium applied the Skills4EOSC methodology to adjust the UPSKILLS course "Introduction to Language Data: Standards and Repositories" within the Humanities and Cultural Heritage Italian Open Science Cloud project. Originally compliant with FAIR principles, the course provides BA/MA linguistics students and instructors with resources and activities on managing linguistic data using certified repositories. CLARIN-IT translated the course into Italian and published it as Markdown files, enhancing its reusability. Selected learning content covering language resources, repositories, and CLARIN core services, like the Virtual Language Observatory, were then tailored for a selection of classes at the University of Ferrara to emphasize the importance of data management in language studies. This pilot applied a train-the-trainer approach and allowed for the participation of researchers and professors from the Linguistic Department of the University of Ferrara, which has been a CLARIN-IT member since 2023.

Outcomes

By the end of this presentation, the workshop participants will learn through a concrete example of how the research infrastructure supports a network of trainers and researchers through reusable learning resources, and especially how the FAIR-by-Design method can be applied to adapt existing resources for other teaching contexts in DHSS/DHCH education.

Bibliography

CLARIN Learning Hub: https://www.clarin.eu/content/learning-hub

Humanities and Cultural Heritage Italian Open Science Cloud project (H2IOSC): https://www.h2iosc.cnr.it/
H2IOSC Training Environment: https://h2iosc-training-platform.ilc4clarin.ilc.cnr.it/login

Filiposka, S., Mishev, A., Kjorveziroski, V., & Leister, C. (2024, July 1). FAIR-by-Design Learning Materials Methodology Training of Trainers. Zenodo. https://doi.org/10.5281/zenodo.12604767

van der Lek, Iulianna; Fišer, Darja. (2023). Introduction to Language Data: Standards and Repositories. In UPSKILLS Learning Content. https://upskillsproject.eu/project/standards_repositories/. CC BY 4.0

Iulianna van der Lek[1], **Anna Woldrich**[2], **Tanja Wissik**[2]
[1]CLARIN ERIC, Netherlands, The; [2]Austrian Centre for Digital Humanities and Cultural Heritage, Austrian Academy of Sciences Vienna, Austria

AI in Higher Education: A CLARIN Community Survey

Wednesday, 05/Mar/2025 4:40pm - 5:00pm
ID: 269 / Session SP 04: 3
Short paper (full-text) | 15-minute presentation with a 5-minute Q&A
Keywords: AI, research infrastructure, higher education, research, teaching, CLARIN

The easy access to AI technologies such as ChatGPT might have huge implications for teaching and learning in higher education. Therefore, more and more policies to regulate the use of AI are published by countries and institutions but there is no general overview of the existing policies in the CLARIN member states and their higher education institutions. There is also no overview of how the CLARIN community is applying AI technologies in the classroom; furthermore, there is no data available on whether the CLARIN community has specific concerns regarding the use of AI in Education and Research. To fill this gap, a CLARIN community survey was launched in early 2024 as a pilot study. Hence, this contribution reports on the results of the survey conducted.

Petrina Vasileiou[1], Jasmine Aavaranta Hansén[1], Gunnar Almevik[2], Elin Fornander[1], William Illsley[3], Wilhelm Lagercrantz[1], Jonathan Westin[2]

[1]National Historical Museums of Sweden, Sweden; [2]Gothenburg University; [3]Swedish National Data Service

Merging innovation with educational impact: XR technologies in the GLAM sector

ID: 317 / Poster Session 2: 34
Late-breaking poster
Keywords: XR technologies, Pedagogy, Museums, Source criticism, Immersive experiences

XR-technologies are becoming increasingly influential in the GLAM-sector, offering new opportunities to enhance traditional educational practices. These technologies, which, among others, include virtual reality (VR), augmented reality (AR), and immersive media, provide museums with novel ways to engage audiences and convey information. However, the potential of XR technologies does not automatically translate into desirable educational outcomes. The challenge lies in ensuring that the experience is not just technically impressive, but also pedagogically meaningful.

Museums have long been trusted institutions for evidence-based knowledge and as they incorporate XR technologies in their exhibitions, it is important to maintain this trust. This includes addressing concerns about source criticism, transparency and credibility when producing new, immersive content. The project aims to produce guides and documents that cultural heritage actors can consult when faced with the pedagogical choices that these technologies offer.

To form an overview of the field we have systematically collected data across Europe on more than seven hundred unique exhibitions with immersive elements, and organised them by key themes related to technology, experience, pedagogical approaches, and aims. This process helped us identify trends and patterns in how XR technologies are being implemented in the museum sector. Four prominent trends emerged from our data:

1. Limited use of VR and AR:
   Despite their popularity, these technologies remain less common in the GLAM sector. Their use is mainly concentrated in major cities and well-established institutions where collaborations with other organisations help provide the technical and cultural capital necessary to implement such technologies/ experiences.

1. Science and Technical Museums show greater adoption:
   These types of museums have been more likely to integrate immersive experiences into their permanent exhibitions compared to historical museums, a fact that underscores a diversity in knowledge and prior experience in the landscape of immersive technologies.

1. Touring projection-based art exhibitions is a growing trend:
   These large-scale art exhibitions utilize projection-based immersive experiences that boast a bombastic format and communication style, which in turn, attracts large crowds and garner significant media attention. Featuring famous artists, they promise audiences a fresh perspective on artworks and appeal to all age-groups.

1. Popularity of 360-degree media and variety in terms of quality:
   360-degree photography and video, along with interactive three-sixty-exhibitions, emerge as the predominant immersive digital initiatives. This technology is accessible to a variety of contributors (from educators, researchers, and established museum institutions, to individuals and businesses of various size) and has a very low threshold. This dominant category includes everything from spontaneous documentation of historical sites to meticulously scanned replicas of entire permanent exhibition floors and archival documentation of past exhibitions.

This poster highlights the importance of aligning the technical aspects of XR experiences with the educational goals of museums. As the field continues to evolve, our research study aims to identify and provide the sector with best practices for using XR technologies in ways that prioritize educational value over novelty, ensuring that they become a welcome addition and continuation of over a century of pedagogical learning and development.

Mo von Bychelberg
Uppsala University, Sweden

Artificial Intelligence Technology’s Influence on the Authenticity of Digital Intangible Cultural Heritage Archives

ID: 136 / S01: 3
Doctoral Consortium
Keywords: Artificial Intelligence, Authenticity, Archives, Heritage, Digitalization

This thesis focuses on the theory and practice of archiving digital intangible cultural heritage (ICH) in the form of martial arts heritage with the help of artificial intelligence (AI), and the consequences this entails for the archive in regards to authenticity of cultural heritage. In just a few years, AI, especially machine learning (ML), has become ubiquitous in vastly different areas of society and culture, including safeguarding cultural heritage. While the preservation of material cultural heritage for future usage with the help of AI, for example using Handwritten Text Recognition (HTR) for making archived manuscripts accessible for readers today as well as searchable, has been practiced for some time, the potential of AI and ML for archiving intangible cultural heritage has been less focused on. This thesis considers implications of using AI and ML for archiving ICH in the form of *digital intangible cultural heritage* – a term I will discuss in the thesis –, with a focus on how preserving ICH with the help of AI influences the authenticity of digital ICH heritage.

For the theoretical framework, I draw on several related theories. Regarding embodied archival heritage, the framework encompasses discussion of embodied information practices (Olsson & Lloyd, 2017), embodied knowledge (Alliata et al., 2024; Craig et al., 2018), the cultural archive (Said, 1994), the human body as an archive (Simas, 2022; Alliata et al., 2024), and the concept of living human treasures (Aikawa-Faure, 2014; Rossi, 2018). In this thesis, the human body is regarded as a carrier of cultural knowledge and memory; it retains elements of cultural expressions such as types of gestures and movements, and has the capability of transmitting this knowledge into the future. As such, the human body can be considered an ‘archive’ in itself, hosting various kinds of embodied ‘records’ that can be passed on to others in the form of performances and teaching. In order to safeguard this embodied heritage for future users, efforts to digitize this type of heritage have been made in the form of various archival projects such as the Hong Kong Martial Arts Living Archive (HKMALA) (Hou et al., 2023), creating archives of digital embodied heritage: a digital version of embodied forms of cultural expression. The embodied ‘records’ are represented digitally and, in some cases, also enhanced via AI in the form of visualization and retrieval (Alliata et al., 2024). Evidently, this digitization process raises questions concerning if, and how, the authenticity of the embodied heritage can be preserved digitally. In this thesis, authenticity as a key concept is discussed in relation to archives, cultural heritage and digitization. The focus here lies on the practices in digital cultural archives preserving digital embodied heritage; what they imply regarding ideas of authentic heritage and how authenticity is perceived and practiced in these archives.

The study is based on semi-structured qualitative interviews with staff and participants at two different digital cultural archives, and a document analysis comprised of official documents relating to the archives. The first case study is the Hong Kong Martial Arts Living Archive (HKMALA), which has been working on preserving martial arts digitally via motion capture. The second project is the Taekwondowon (National Taekwondo Center) of Korea, which to my knowledge does not work with AI techniques. The findings of these two complementary studies are then compared with the affordances of state-of-the-art AI techniques available for preserving, processing, and passing on digital embodied heritage. The outcome of this comparison is an AI conceptual framework suggesting a possible AI-supported preservation strategy for ICH in the form of digital embodied heritage.

The thesis focuses on the following research questions:

RQ1: How do archivists theorize and practice the archiving of digital embodied heritage?

RQ2: How do archivists theorize and practice the safeguarding of authenticity in archiving digital embodied heritage?

RQ3: What are the implications of using AI to archive digital embodied heritage for the authenticity of the archive?

Bibliography

Aikawa-Faure, N. (2014). Excellence and authenticity: ‘Living National (Human) Treasures’ in Japan and Korea. International Journal of Intangible Heritage, 9, 37-51.

Alliata, G., Hou, Y., & Kenderdine, S. (2024). Augmenting Access to Embodied Knowledge Archives: A Computational Framework. Digital Humanities Quarterly, 18(3).

Craig, C. J., You, J., Zou, Y., Verma, R., Stokes, D., Evans, P., & Curtis, G. (2018). The embodied nature of narrative knowledge: A cross-study analysis of embodied knowledge in teaching, learning, and life. Teaching and Teacher Education, 71, 329-340.

Hou, Y., Seydou, F. M., & Kenderdine, S. (2023). Unlocking a multimodal archive of Southern Chinese martial arts through embodied cues. Journal of Documentation, ahead-of-print. https://doi.org/10.1108/JD-01-2022-0027

Olsson, M., & Lloyd, A. (2017). Being in place: embodied information practices. Information Research, 22(1), CoLIS paper 1601.

Rossi, J. (2018). The “Living Human Treasures” System in the Republic of Korea [Master’s Thesis, Università Ca’Foscari Venezia]. Archivio istituzionale ad accesso aperto. http://hdl.handle.net/10579/13069

Said, E. W. (1994). Culture and imperialism. Vintage.

Ruilin Wang[1], **Lidia Pivovarova**[1], **Yann Ryan**[2], **Mikko Tolonen**[1]
[1]University of Helsinki; [2]Leiden University

Image-reuse Identification in a Large Collection of Eighteenth-century British Books

ID: 139 / Poster Session 2: 7
Poster and demo (abstract) with accompanying a 1-minute lightning talk
Keywords: historical book analysis, self-supervised representation learning, headpiece, Tonson, publisher networks

Historical books convey substantial information through visual elements, including illustrations and ornaments. In the handpress era, these headpieces were impressed onto the page with ink using reusable carved or engraved templates, resulting in repeated copies of certain woodcuts. Woodcuts were more expensive than text types, and thus were used in many different books. In this paper we deal with automatic grouping of book images, considering both identical images and near-duplicates. When performed on a large scale and combined with metadata, this analysis sheds light on image circulation and publishing culture in the past. We demonstrate how such analysis can be used to study books published by the Tonson family, a prominent dynasty of publishers in the eighteenth century, and to uncover more complex printing practices of the eighteenth century than commonly understood.

Approach

Our study focuses on headpieces—ornaments used at the top of the page at the beginning of a chapter—extracted from Eighteenth Century Collections Online (ECCO), a collection of approximately 200,000 digitized 18th-century books. We use self-supervised contrastive representation learning and clustering to group these images, revealing those that might be created from the same woodblock. Manual annotation is then used to clean up the clusters, and the resulting dataset is used to fine-tune the model. This semi-supervised approach yields 96% average precision in identifying similar headpieces. Fine-grained clusters are then organised in larger groups of near-duplicates.

As a result of this process we find large groups of similar images, up to several hundred headpieces within a group. For some previously studied images we found by an order of magnitude more examples that were known before. That is more, large-scale data-driven analysis reveals interesting patterns of image circulation between publishing houses and cities. We discuss one such case in the next section.

Analysis

Our analysis reveals intriguing patterns in the use of headpiece ornaments associated with the Tonson-Watts publishing enterprise, where Watts was a printer working for Tonson but also publishing some books on his own. Figure 1 shows how selected headpieces prominent in Tonson’s books were used between 1720 and 1739. We observed a sharp decline in the use of these headpieces after Jacob Tonson Jr.'s death in 1735, despite John Watts' continued printing activities. This suggests that not all Tonson-Watts ornaments remained in Watts' stock, highlighting the complex ownership dynamics of publishing tools. We also found a correlation between the Tonson empire's decline and the overall decrease in headpiece use, which underscores the Tonson publishing house's influence in the 1720s and 1730s.

A fascinating case of ornament reuse involves the "angel_head_trumpet_birds" (C002) design, shown in Figure 2. We identified distinct versions of this ornament: one associated with Tonson/Watts prints in London (C002_01) and another with Irish printing (C002_02). The deliberate differentiation of these headpieces, coupled with their sustained use over 30 years, suggests a strategic use of ornaments beyond mere imitation. This distinction raises questions about Tonson's possible endorsement of Irish editions, challenging the previous assumption of a hard one-to-one connection between an image and a publisher.

Rebecka Weegar[1], Kajsa Palm[1], Roger Mähler[1], Fredrik Mohammadi Norén[2], Johan Jarlbrink[1]
[1]Umeå University, Sweden; [2]Malmö University, Sweden

SweDeb: An Interface to Explore Swedish Parliamentary Debates Since 1867

ID: 143 / Poster Session 2: 1
Poster and demo (abstract) with accompanying a 1-minute lightning talk
Keywords: Parliamentary debates, Open data, Natural language processing

The SweDeb project aims to create a user friendly and openly available interface for examining parliamentary data from Sweden. The SweDeb interface makes it possible to explore and analyse the content of parliamentary debates from 1867 and onwards using tools for filtering and natural language processing. This poster will present a first version of the interface.

The corpus used with the SweDeb interface is based on openly available data from the Swedish parliament. The corpus has been further processed within the projects WeStAc and SWERIK, which have re-OCRed the debate records, segmented the texts into individual speeches and enriched the corpus with metadata, with continuous improvements of the curated quality. In total, the corpus contains over 1 million annotated speeches, comprising some 450 million speech words. The estimated mapping between speaking members of parliament and their respective annotated speeches is about 87 % over the whole time period but often above 95 % after 1920 (Yrjänäinen et al., 2024).

This rich dataset, documenting a vast number of political debates in Sweden, has many use cases and users, both for researchers and students within digital humanities and in other fields – history, political science, sociology et cetera – and for journalists, officials and politicians, as well as for members of the public. The creation of an accessible and easy-to-use interface makes this corpus readily available to a larger audience, and the addition of metadata makes it possible, through the SweDeb interface, to search for speeches by speaker name, party, gender, chamber and time period, and to download custom datasets.

The SweDeb interface offers several tools for searching and analysing the corpus. In addition to filtering the dataset based on metadata criteria, it is also possible to search for words and examine how they have been used over time and by different groups, to get the context of a search term or phrase, and to search for n-grams. The SweDeb interface is also intended to be extended with additional tools, such as topic modelling, and with other data sources (e.g. the motions), further improving access to the content and language of parliamentary texts for digital huminites-oriented scholars and beyond.

Bibliography

Väinö Yrjänäinen, Fredrik Mohammadi Norén, Robert Borges, Johan Jarlbrink, Lotta Åberg Brorsson, Anders P Olsson, Pelle Snickars, Måns Magnusson. 2024. “The Swedish Parliament Corpus 1867–2022”. In Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), 16100-16112.

SweDeb, https://swedeb.se/public/index.html#/

Jonathan K. Westin, **Gunnar Almevik**
University of Gothenburg, Sweden

The Inscriptions of Saint Sophia in Kyiv - Participatory and research grounded approaches to data collection during war

Thursday, 06/Mar/2025 3:40pm - 4:00pm
ID: 245 / Session SP 08: 6
Short paper (abstract) | 15-minute presentation with a 5-minute Q&A
Keywords: Stakeholders, on-site-knowledge, data collection

This short paper focuses on community engagement, participatory approaches and inclusion in digital humanities initiatives as a methodology and guiding principle for data collection.

Carved into the walls of Saint Sophia Cathedral in Kyiv, Ukraine, are more than 7,000 inscriptions that span over a thousand years, using different language and writing systems (both ancient and modern). The inscriptions constitute not only an incomparable source of knowledge about the cultural history of man, language and writing, migration, cultural exchange, and diversity, but also a most challenging material to document and study: these inscriptions are multilingual, multimodal, and multi-temporal as well as layered upon one another with successive layers of messages etched over earlier ones during the centuries. Furthermore, the inscriptions are a precious source material for historical and cultural research that, due to Russia's warfare and bombing, is under dire threat. Since March 2022 more than 460 culturally significant environments and monuments have been destroyed (UNESCO 2025), and though Saint Sophia Cathedral is not considered a target there is an awareness that due to its location close to vital infrastructure it is threatened by drone and missile attacks. This not only prevents the international community of researchers from visiting the cathedral during the war, but if the cathedral is severely damaged the digitised material will be one of the only available sources for future studies of the inscriptions. To safeguard this cultural heritage and research source, an ongoing collaboration between Ukraine and Sweden are documenting the inscriptions and publishing the data through an online portal that allows for both a manual visual analysis of the digitised surfaces and inscriptions but also an infrastructure for data registration and computational analysis using artificial intelligence (Westin *et al.* 2024).

In an effort to meet these challenges and carry out a project in a country burdened by war, the Department of Conservation and the Gothenburg Research Infrastructure in Digital Humanities (GRIDH) at the University of Gothenburg have worked with stakeholder interviews and knowledge transfer and participatory approaches to data collection. The method, partly developed within the research project *Methods for Digital Diagnosis of Threatened Cultural Heritage* (Westin and Almevik 2023), serves to highlight and counter two conventions within the heritage sector. The first one concerns how the concepts *documentation* and *digitisation* are often uncritically thought to be final and complete acts. Hence, a church can be described as having been documented and an artefact as having been digitized without necessarily prompting any questions about the scope of the documentation or digitisation. By instead framing these as *data collection processes*, one opens from the start up for questions about what aspects are to be collected and how the data should be analysed or used. It creates an awareness that this is only a sampling of a delimited part of the monument, artefact or phenomenon studied. The second one concerns the dichotomy between heritage professional and technical expertise. All too often the technical knowledge is not shared with the heritage professionals but is instead owned by an outside contractor or expert group. The heritage professionals thereby lose control of the digitisation process when it comes to methods, access to the collected data, and storage solutions.

With these experiences as a stepping stone, the project *Digital documentation of inscriptions in the Saint Sophia Cathedral in Kyiv* was instigated by an investigation into who the stakeholders were and what they would require from the collected data to meet their different needs. This involved both deep interviews with individual researchers and forming a reference group of stakeholders. The reference group served also as a basis for reaching an even larger and more diverse group of experts to participate in a first workshop and evaluation of the proposed digitisation methods and their result. This evaluation served to help the project further develop the methods and do a course correct to meet the demands of the stakeholders. The personnel at the Saint Sophia Cathedral Museum have served an integral role in the process of testing and evaluating digitisation methods on site and have then been trained both in the necessary workflows to independently carry out high quality digitisations of inscriptions, preprocess and evaluate the collected data, and to employ proper data management.

Bibliography

UNESCO (2025). Damaged cultural sites in Ukraine verified by UNESCO. https://www.unesco.org/en/articles/damaged-cultural-sites-ukraine-verified-unesco Accessed 2025.01.23

Westin, Jonathan and Gunnar Almevik (2024). Metoder för digital diagnostisering av hotade kulturarv (RAÄ2021-2648). Riksantikvarieämbetet.

Westin, Jonathan, Gunnar Almevik, Tristan Bridge, Matteo Tomasini and Ashely Green (2024). *Saint Sophia's Inscriptions*. v.1.1 Gothenburg Research Infrastructure in Digital Humanities. https://saintsophia.dh.gu.se/. Accessed 2025.01.23

Jonathan K. Westin, **Daniel Brodén, Mats Fridlund**
University of Gothenburg, Sweden

Augmenting digital publications: From text and image to rich visualisations of cultural heritage data

Thursday, 06/Mar/2025 10:30am - 11:00am
ID: 242 / Session LP 04: 4
Long paper (abstract) | 20-minute presentation with a 10-minute Q&A
Keywords: Digital publication, academic communication, NTRO, multi-modal

1. Introduction

This paper addresses how digital portals, tools, and datasets can be effectively developed and evaluated as legitimate non-traditional research outputs (NTROs) within the academic merit system. It also explores strategies to enhance the scholarly recognition, authorship, and long-term accessibility of digital publications in humanities and social science research, grounded in the work at GRIDH. To further shed light on the state of digital publications within academia, the aim of the present paper is to collate and present a broad set of interdisciplinary experiences of researchers, research engineers and collaborative partners at GRIDH, focusing on structuring data for rich spatio-temporal visualisations.

We begin by highlighting the problem of relying on textual outputs in academia, contrasting traditional practices with the potential of new digital formats for improved knowledge transfer. The paper proceeds by outlining GRIDH's approach to facilitating interdisciplinary projects and collaborations between technical experts and humanities researchers to meet the evolving standards in digital humanities and represent the ways in which historical artifacts generated meaning in the past. Then, we provide concrete examples of GRIDH projects and technical strategies for presenting data-intensive research interactively that demonstrate expertise in structuring data for rich spatio-temporal visualisations, that manifests the need to reduce the translation between analysis and dissemination. We conclude by emphasising the benefits of integrating NTROs into academic evaluation systems, addressing issues of authorship and the longevity of digital formats.

2. A notable gap

It has been argued that the academic system remains heavily reliant on textual output (see Berry & Fagerjord 2017). The conventional academic process centers on producing written content across a variety of textual formats, including full research articles, conference papers, technical reports, case studies, reviews, books, and research proposals (Almevik and Westin 2022). To conform to these genres, research results need to be translated into words, a process further disciplined by the accepted disposition and rules of the format. However, it has also been argued that while this conforms the research into formats that are easy to review and share, it does a disservice to the knowledge transfer and the transparency when it comes to the presentation of the data. New technologies today provide a diverse array of formats that can significantly improve research communication and reduce information loss when translating between different modes, media, and formats. The latest advancements in informatics, digital humanities, and multimodal anthropology investigate how gaming, social networks, and immersive or augmented reality are not only transforming societal practices but also redefining the practice of research itself (Gubrium, Harper and Otanez 2015; Collins, Durington and Harjant 2017; Almevik and Westin 2022).

While substantial research has explored multimodal methods for data collection and analysis – such as using film to reveal skills or embodied knowledge, or employing 3D representations to investigate materiality – there is a notable gap in addressing how these multimodal approaches and emerging technologies can be effectively integrated into the final research outputs. Moreover, as Gunnar Almevik and Jonathan Westin argue, there is a need to balance these innovations with the dual demands of academic rigour and practical relevance (2022). In the context of artistic research, augmented tools are often crucial for capturing the subtle nuances of both processes and qualities in practice. For instance, film may be indispensable for illustrating variations in motion or supporting the analysis of sensory experiences, while a detailed 3D model of an object, material, or structure may be vital for outlining the research focus or understanding the intricate relationships between different surfaces. This is evidenced by the ample research on the technical aspects of new media and how digital technologies can be utilised to communicate research (see Debevec 2005; Kahr-Højland 2007). In the context of academic work and publishing, 3D documentation, sound, video and code examples thus have several advantages as they are providing a framework where less abstraction is needed to establish knowledge transfer and evaluation of ideas.

3. Integrative interdisciplinary collaborations

Since 2015, GRIDH (formerly the Centre for Digital Humanities, CDH) has initiated and developed data-rich research projects and digital resources (tools, databases, archives, etc.) in collaboration with researchers from the humanities and other faculties. The work at GRIDH is grounded in what we describe as interdisciplinary project design (Brodén et al. 2024), which refers to a specific approach for integrating collaboration between technical experts and 'traditional' humanities and social science researchers. This approach addresses multidisciplinary teamwork requirements and also feeds into the ongoing discussion in digital humanities about the need for context-sensitive approaches that are based on domain knowledge. Critical commentators in the field emphasise the importance of engaging with the original context of archival data to produce robust and nuanced results. Katherine Bode (2018) critiques the tendency to rely on data models that inadequately represent the ways in which historical artifacts generated meaning in the past, while Jo Guldi (2023) cautions against naïve assumptions about the relationship between data and the documentary record. We find the notion of interdisciplinary project design useful for delineating an emergent form of expertise in the 'fractured trading zones of digital humanities' (Svensson 2011). On some level, this expertise resembles what Andrea Hunter (2014: 27) refers to as 'bridge people', who can speak the languages of 'the two cultures' and acknowledge domain specific needs and contexts.

4. Reducing the translation between analysis and dissemination

With a start in the projects *Mapping Lived Religion* (2019–2024), *Expansion and Diversity* (2019–2021) and *Pehr Strands Flöjtur* (2018–2021), GRIDH developed expertise in structuring data for rich spatio-temporal visualisations. This expertise has been further developed in subsequent projects through the development of a package of frontend modules organised around a data model specifically aimed at cultural heritage data and the design of three distinct interaction models: one built around a traditional paper structure where the user advances through various steps akin to individual chapters and the interactive exploration of data is just one of these; one arranged as a series of distinct and interconnected views that lets the user interact with the data at different scales and levels of abstraction, and one where the user explores the data through a series of tools specifically created to study certain aspects or sections of the dataset. Through a number of projects such as *Extended Rephotography* (2020–

2024)*, Reading the Signs* (2022–2024)*, Göteborgs Jubileum 1923* (2023)*, Etruscan Chamber Tombs* (2023-2024)*, Svenskt Digitalt Orgelarkiv* (2020–2024)*, Saint Sophia's Inscriptions* (2024–2026)*, Maritime Encounters* (2022–2027), and *New Paths to the Past* (2020–), these interaction models have been expanded with capabilities to view, perform measurements on, and evaluate 3D data, explore reflectance transformation imaging (RTI), browse and filter visual galleries of datasets, and group and sort documentation according to date or type. Hence, the interaction models are defined by a 'linear modularity', a semi-rigid structure that moves from the textual or visual establishment of context to the exploration of published data through media-specific tools. However, a key differentiator between traditional and digital research output lies in the relation and access to underlying data either as downloadable datasets or open API's that allows others to incorporate the data in their research or projects. Hence for its backend, GRIDH's publication strategy relies on a database coordination solution written in Django with PostgreSQL. The backend solution allows for the serialisation and generation of generic and consistent views in the form of representational state transfer APIs, through the Django REST framework. This ensures the creation of compliant web APIs that both the frontend of the publication relies upon, and that can be published as open APIs (see Westin, Bridge and Tomasini 2024).

While NTROs have a clear head start in artistic research they are beginning to earn wider academic recognition. Research councils and national assessment bodies have included NTROs in guidelines for assessment of research for more than a decade (ARC 2014; University of Sidney 2014; Barwick and Toltz 2017). However, while the digital publications GRIDH have developed together with various research groups all present research in interactive and engaging ways that reduces the translation between analysis and dissemination, outside artistic research they are seldom upheld by the involved researchers as publications in their own right, nor are they incorporated in university merit systems. This might be traced back to several circumstances, both in how the publications are created as a multi-author and trans-disciplinary collaboration between researchers and research engineers with sometimes unclear authorships and transparency when it comes to individual contributions, and as an effect of a systematic mistrust in the longevity of digital publications.

5. Conclusions

In conclusion, the efforts at GRIDH to develop digital portals, tools, and datasets highlight the potential of non-traditional research outputs (NTROs) to enrich academic communication and research dissemination, particularly through data visualisation and open access to datasets. While NTROs are gaining recognition, especially within artistic research, there remains a need to better integrate these digital publications into academic evaluation systems, addressing issues related to authorship, collaboration, and the perceived longevity of digital formats.

Bibliography

Almevik, Gunnar, and Jonathan Westin (2022). "Rethinking the Academic Artefacts". In Craft Sciences, edited by Tina Westerlund, Camilla Groth and Gunnar Almevik. Gothenburg: Kriterium, Acta Studies in Conservation.

Australian Research Council (ARC). 2014. Excellence in Research for Australia: ERA 2015, Submission Guidelines. Canberra: Commonwealth of Australia.

Barwick, Linda, and Joseph Toltz. 2017. "Quantifying the Ineffable? The University of Sydney's 2014 Guidelines for Non-Traditional Research Outputs." In Perspectives on Artistic Research in Music, edited by Robert Burke and Andrys Onsman, 67–77. Oxford: Lexington Books

Berry, David, Ander Fagerjord (2017): Digital humanities, Polity.

Bode, Katherine (2018): A world of fiction. Digital collections and the future of literary history, University of Michigan Press.

Brodén, Daniel, Mats Fridlund, Cecilia Lindhé and Jonathan Westin (2024): 'Designing digitally-driven interdisciplinarity: Between protocol and judgement', Proceedings of the Huminfra Conference (HiC 2024), Linköping Electronic Conference Proceedings 205, 128–134.

Collins, Samuel, Matthew Durington, and Gill Harjant (2017). "Multimodality: An Invitation." American Anthropologist 119 (1): 142–46.

Debevec, Paul (2005). "Making 'The Parthenon'." 6th International Symposium on Virtual Reality, Archaeology, and Cultural Heritage, Proc. VAST 2005, Pisa, Italy.

Gubrium, Aline, Krista Harper, and Marty Otanez (2015). Participatory Visual and Digital Research in Action. Walnut Creek: Left Coast Press.

Guldi, Jo (2023): The dangerous art of text mining: A methodology for digital history, Cambridge University Press.

Hunter, Andrea (2014): 'Digital Humanities as third culture', MedieKultur, 57, 18–33.

Kahr-Højland, A. 2007. "Brave New World: Mobile Phones, Museums and Learning." The Journal of Nordic Museology (1): 3–19.

Svensson, Patrik (2011): 'The Digital Humanities as a Humanities Project', In M Gorman (Ed.): Trading zones and interactional expertise: Creating new kinds of collaboration, MIT Press: 42–60.

University of Sydney Research Portfolio. 2014. University Guidelines for Non-Traditional Research Outputs (NTROs). Sydney: The University of Sydney.

Westin, Jonathan, Tristan Bridge & Matteo Tomasini (2024), From the Arctics to Antarctica - A multimodular visualisation of data, Proceedings of the Huminfra Conference (HiC 2024).

Johannes Widegren
Linnaeus University, Sweden

Automatic subject indexing of oral history interviews with Whisper and ChatGPT

ID: 114 / Poster Session 2: 18
Poster and demo (abstract) with accompanying a 1-minute lightning talk
Keywords: Artificial intelligence, oral history, automatic transcription, automatic subject indexing

In the archival media trinity of text, image and sound, the latter presents particular challenges for users' searching and browsing. While a user can manually or digitally browse through texts and images to locate items of interest, sound needs to be played more or less in real time to do the same. Audio files can naturally be described and transcribed – digital or digitized audio files of speech automatically so – thus facilitating full-text search. However, the familiar Achilles' heel of full-text search, namely the ambiguity of natural language, remains.

Enter the hallmark trade of librarianship: subject indexing. When subject index terms have been assigned to individual sections of an audio file, a user searching for these or similar terms can locate precisely where in the retrieved audio files the subject is discussed. With state-of-the-art AI systems, even this can be done automatically, thereby decimating the amount of time needed to index audio files, from real time to as fast as the system can process them. This heralds a brave new future for the accessibility and searchability of oral history archives.

This poster presents a pilot study on automatically transcribing interviews in Swedish from oral history archives using OpenAI's Whisper, describing the content and assigning subject index terms to sections using OpenAI's ChatGPT, and visualizing the results. The accuracy of the results depends on many factors, including sound quality, accents of the speakers, the amount of language mixing etc. The results are very promising, however, suggesting automatic subject indexing of interviews to be a worthwhile research direction going forward.

Johannes Widegren
Linnaeus University, Sweden

Human-centered AI approaches for improved information discoverability in Sámi archival collections

ID: 115 / **S01: 2**
Doctoral Consortium
Keywords: artificial intelligence, archives, sámi archives

Postmodern archival scholarship has revealed the power dynamics inherent in the shaping of collective memory through archives and recordkeeping. Rather than constituting neutral evidence of past events, archives are deliberately formed by dominant groups in society to assert power over marginalized groups (Dunbar, 2006). Western hegemonic notions of what constitutes a record exclude the narratives of e.g. Indigenous populations with oral traditions from the writing of history, leading to biased representations of the past that masquerade as the only truth.

In scholarly circles, decades of research have rendered Sweden's roughly 400-year-history of colonial oppression of the Indigenous Sámi populating northern Fennoscandia well known. In spite of this, public and political awareness of the fact has lagged behind, and Sweden has long maintained an air of innocence in its discourse on the matter (Fur, 2016). This has become painfully apparent in recent years, during which Swedish court rulings and policy decisions have failed to recognize the rights of the Sámi as an Indigenous people, inducing international critique. Promoting Sámi counternarratives, to complement or contest the dominant Swedish narrative, has thus become not only a goal to benefit future generations but an endeavor with political ramifications in the now.

Digitalization of archives presents an opportunity for remediating the colonial dynamics of the physical cultural record in line with Risam's (2019) framing of postcolonial digital humanities. A current digitalization project by the Sámi periodical *Samefolket*, presumably the oldest Indigenous periodical in the world, aims to do just that by facilitating access to their archive of over a century of Sámi voices for a wider audience. Having functioned as a vehicle of Sámi opinion, a unifying organizational force and a political platform, this archive is an invaluable source for researchers and the public alike.

The digitalization project team is small, however, and resources are limited. Structuring and adding metadata to digital collections to facilitate information discovery can be a monumental task. For this reason, many cultural heritage institutions have experimented with approaches using artificial intelligence (AI) to structure, describe and create access to digital collections in short periods of time (see e.g. Carter et al., 2022; Luthra et al., 2023). There are inherent risks in using AI technologies, however, which include but are not limited to biased models, data security risks and inaccuracies in output (Foka et al., 2023). In sensitive cultural settings, these technologies should preferably be used in human-AI collaborative workflows.

This project aims to work together with the digitalization team at *Samefolket* to explore the risks and possibilities of using AI to assist the endeavor of creating digital access to their archive, which in addition to the issues of the periodical, written in Swedish and occasionally Sámi languages, contains over 50,000 photos and numerous administrative records and other source material. Which technologies to test will be decided in collaboration with the digitalization team, in order to facilitate their work and strive towards their goals. Potential applications include, in addition to optical character recognition (OCR), topic modeling, named entity recognition, image tagging, semantic clustering of images, text summarization, and more. Thereby it seeks to answer the question of if and how AI can help raise social consciousness of Sámi culture and history, and future-proof Sámi cultural heritage.

Bibliography

Carter, K. S., Gondek, A., Underwood, W., Randby, T., & Marciano, R. (2022). Using AI and ML to optimize information discovery in under-utilized, Holocaust-related records. AI & SOCIETY, 37(3), 837–858. https://doi.org/10.1007/s00146-021-01368-w

Dunbar, A. W. (2006). Introducing critical race theory to archival discourse: Getting the conversation started. Archival Science, 6(1), 109–129. https://doi.org/10.1007/s10502-006-9022-6

Foka, A., Eklund, L., Sundnes Løvlie, A., & Griffin, G. (2023). Critically assessing AI/ML for cultural heritage: Potentials and challenges. In S. Lindgren (Ed.), Handbook of Critical Studies of Artificial Intelligence (pp. 815–825). Edward Elgar Publishing. https://doi.org/10.4337/9781803928562.00082

Fur, G. (2016). Colonial fantasies – American Indians, indigenous peoples, and a Swedish discourse of innocence. National Identities, 18(1), 11–33. https://doi.org/10.1080/14608944.2016.1095489

Luthra, M., Todorov, K., Jeurgens, C., & Colavizza, G. (2023). Unsilencing colonial archives via automated entity recognition. Journal of Documentation. https://doi.org/10.1108/JD-02-2022-0038

Risam, R. (2019). New digital worlds: Postcolonial digital humanities in theory, praxis, and pedagogy. Northwestern University Press.

Johannes Widegren
Linnaeus University, Sweden

Automatic subject indexing of Sámi oral history interviews with Whisper and ChatGPT

ID: 285 / **WS10: 6**
Tradition Archives Meet Digital Humanities II
Keywords: Sámi culture, artificial intelligence, oral history, automatic transcription, automatic subject indexing

In the archival media trinity of text, image and sound, the latter presents particular challenges for users' searching and browsing. While a user can manually or digitally browse through texts and images to locate items of interest, sound needs to be played more or less in real time to do the same. Audio files can naturally be described and transcribed – digital or digitized audio files of speech automatically so – thus facilitating full-text search. However, the familiar Achilles' heel of full-text search, namely the ambiguity of natural language, remains.

Enter the hallmark trade of librarianship: subject indexing. When subject index terms have been assigned to individual sections of an audio file, a user searching for these or similar terms can locate precisely where in the retrieved audio files the subject is discussed. With state-of-the-art AI systems, even this can be done automatically, thereby decimating the amount of time needed to index audio files, from real time to as fast as the system can process them. This heralds a brave new future for the accessibility and searchability of oral history archives.

These methods favor particular kinds of material, however, notably monolingual data in major languages. NLP technologies offer much less support for small languages such as the different varieties of Sámi spoken in northern Fennoscandia. Rich collections of oral history interviews both in Sámi languages and in Swedish with Sámi content are available online, some with little or no finding aids, for which subject indexes would be of immense value. If even basic indexes could be provided automatically, it would increase findability in this culturally vital material manyfold.

This poster presents a pilot study on automatically transcribing interviews in Swedish mixed with Sámi from oral history archives using OpenAI's Whisper, describing the content and assigning subject index terms to sections using OpenAI's ChatGPT, and visualizing the results. The accuracy of the results depends on many factors, including sound quality, accents of the speakers, the amount of language mixing etc. Ethical questions naturally arise when applying AI technologies to Indigenous material, in particular questions of representation and generalization. The results are very promising, however, suggesting automatic subject indexing of interviews to be a worthwhile research direction going forward.

Yunting Xie
Department of Business Studies, Uppsala University, Sweden

From Invention to Innovation: Natural Language Processing to Study Swedish Historical Patents (1890—1945)

ID: 221 / **S01: 6**
Doctoral Consortium
Keywords: patents, NLP, patent value, innovation, Sweden

A patent is an exclusive right granted to inventor for a certain period of time to prevent others from profiting from inventions and their application. In most modern patent systems, examination procedures have been established to evaluate if inventions applied are novel enough to be patented. The purposes for the patent system are to offer the inventor rights to decide on the invention, to acknowledge individual creativity and originality, and to disclose technological knowledge to the public. While initially designed as a national system, the turn of nineteenth and twentieth centuries witnesses the globalization of patenting activities, marked by the Paris Convention in 1883 (Pretel, 2018). The patent has also become an important channel for transnational technology transfer. Arguably, modern patent systems has created both private value and societal value out of innovation activities (Akkoyun et al., 2022). In this paper, the aim is to use natural language processing methods to study the value of patents.

Though not all inventions have been patented, patent data offers concrete evidence for technological innovations and intellectual progress and thus is used widely as the proxy for innovation activities (Ace et al., 2002; Kogan et al. 2017). Particularly for historical periods, patents preserve valuable sources for innovation across different spans, regions, industries and sectors. Most studies using patent statistics quantitatively to measure the output of innovation activities. Yet a major challenge remains, to evaluate the patents more precisely, and to identify important and valuable patents from others, due to a large number of patents granted each year and the vast variation of the quality among them. A key theme in patent history is therefore the value in patenting and innovation: what is value in innovation, and accordingly, how can patents be used to assess such value?

The division between private and social values of patents sheds different lights on the methods for evaluation. The private value consider how valuable a patent is from patentees' view, and the social value regards the valuation by others. The private value has been accessed mainly by yearly fee payments (Schankerman & Pakes, 1986). The more patentees invested for extending the life of their patents, the higher expectations they had for the value. The similar goes for measuring the geographical breadth of patents by constructing historical patent families (Andersson, La Mela & Tell, 2024). However, two patents with similar time and space scope might still differ hugely in their relative values, not only because that patents are designed as a trade-off between interest exclusiveness and information disclosure and thus have a limited lifespan, but also since patentees might have different criteria and strategies to decide whether or not to prolong their patents.

The perspective of social value offers an approach to compare in such scenarios. However, it has been particularly a difficulty to study historical patents, which, unlike the modern ones, have no reference system such as citations to indicate the relative importance before the 1950s. The social value of historical patents are mostly assessed retrospectively using citations from later periods (Nicolas, 2011), or externally with other contemporary sources like biographical dictionaries (Nuvolari & Tartari, 2011; Nuvolari, Tartari & Tranchero, 2021). There has long been a lack of methods for patent evaluation with no extra proxy that might causes turbulence especially when creating longterm indices. This paper addresses the challenge by constructing a text indicator to measure the value of patents granted in Sweden in the period 1890—1945.

Textual indicators on patent specifications are recently developed to measure patent value. The patent publications carries detailed information on the description of inventions, the claims of novelty and the drawings for examiners and later for the public. They thus become a good source to build internal indicators for patent value. The key idea is to measure the similarity between patent specifications: an important patent is distinct from the previous patents in language and influences the language used in the subsequent patents. Various methods of natural language processing has been applied to investigate the similarity, such as new words and bigrams (Arts, Hou & Gomez, 2020), high dimensional textual analysis (Kelly et al., 2021, on US patents from 1840 to 2010; La Mela, Frankemölle & Tell, forthcoming, on Swedish patents from 1890 to 1929), and text embedding (Hain et al., 2022). Using Swedish historical patent in the first half of twentieth century, this paper pushes forward the frontier by studying how language difference impact the validity of text indicators, as well as creating and validating indices for innovation breakthroughs during a period witnessing Sweden's growth in innovation activities at the global stage.

In the study of modern patents, text indicators are more effective method than patent citation (Arts et al., 2020). It is also regarded as more likely to capture the scientific value of a patent instead of the economic one. To contribute to current discussion in patent value, this paper investigates further which aspect of social value is more likely measured by the text indicator in historical studies by triangulation with other primary sources such as biographical dictionaries and industrial periodicals. The paper also investigates how historical technology vocabulary change and morphological differences (e.g. between Swedish and English) may affect the performance of the established text similarity measurements. For the history of innovation in the context of twentieth-century Sweden, this paper improves the indicator constructed on Swedish corpus in a pilot study (La Mela, Frankemölle & Tell, forthcoming) by extending the time scope until World War II, using clips of the documents to minimize the influence of length variation, analyzing the evaluation at the industry and sector level and validating the new indicator with other measures including annual fee payment and industrial references.

Bibliography

Acs, Z. J., Anselin, L., & Varga, A. (2002). Patents and innovation counts as measures of regional production of new knowledge.Research policy,31(7), 1069-1085.

Akkoyun, H. C., M. Andrews, & R. Vasu. (2022). Private patent values: Evidence from a 1919 Illinois Blue Sky Law. Working paper (July 17, 2022) presented at the Paris WEHC 2022 (25-29 July 2022).

Andersson, D., La Mela, M., & Tell, F. (2024). Family first: Defining, constructing, and applying historical patent families.Explorations in economic history, 101627.

Arts, S., Hou, J. & Gomez, J.C. (2021). Natural language processing to identify the creation and impact of new technologies in patent text: Code, data, and new measures. Research Policy, 50(2).

Hain, D. S., R. Jurowetzki, T. Buchmann, & P. Wolf. (2022). A text-embedding-based approach to measuring patent-to-patent technological similarity. Technological Forecasting and Social Change, 177, 121559.

Kelly, B., D. Papanikolaou, A. Seru, & M. Taddy. (2021). Measuring Technological Innovation over the Long Run. American Economic Review: Insights, 3(3), 303–320.

Kogan, L., Papanikolaou, D., Seru, A., & Stoffman, N. (2017). Technological innovation, resource allocation, and growth. Quarterly Journal of Economics, 131 (2) 665–712.

La Mela, M., J. Frankemölle, & F. Tell. (forthcoming). Using document similarity in historical patents to study value in innovation. Proceedings for Digital Humanities in the Baltic and Nordic Countries 8th Conference, May 27-31, 2024, Reykjavík.
Nicholas, T., 2011. Independent invention during the rise of the corporate economy in Britain and Japan. The Economic History Review, 64(3), 995–1023.

Nuvolari, A. & V. Tartari. (2011). Bennet Woodcroft and the value of English patents, 1617–1841. Explorations in Economic History, 48(1), 97–115.

Nuvolari, A., V. Tartari, & M. Tranchero. (2021). Patterns of innovation during the industrial revolution: A reappraisal using a composite indicator of patent quality. Explorations in Economic History, 82, 101419.

Pretel, D. (2018). The global rise of patent expertise during the late nineteenth century.Technology and globalisation: Networks of experts in world history, 129-157.

Schankerman, M. & A. Pakes. (1986). Estimates of the value of patent rights in European countries during the post-1950 period. The Economic Journal, 96(384), 1052–1076.

Yunting Xie[1], **Matti La Mela**[2], **Fredrik Tell**[1]
[1]Department of Business Studies, Uppsala University; [2]Department of ALM, Uppsala University

Multimodal LLM-assisted Information Extraction from Historical Documents: The Case of Swedish Patent Cards (1945-1975) and ChatGPT

Thursday, 06/Mar/2025 10:30am - 11:00am
ID: 257 / Session LP 06: 4
Long paper (full-text) | 20-minute presentation with a 10-minute Q&A
Keywords: information extraction, AI, historical documents, patent history, error flagging

This paper presents an AI-assisted method for information extraction from historical documents using multimodal large language model (MLLM). We develop a pipeline to retrieve text and events from Swedish historical patent cards using the GPT-4o model to extend the Swedish historical patent database. Our study demonstrates how generic MLLMs can help to save time and labor cost for creating applicable data in a low-source setting, which is a common challenge for digital humanities projects leveraging the latest AI technologies. We also explore the error flagging for automated text recognition that can integrate into traditional information extraction workflow: the MLLMs' vision capacity helps to identify documents with potential errors that require human verification. We conclude that the model generates usable yet imperfect data which speeds up data collection and reduces its cost. The flags created simultaneously in information extraction enable to evaluate the model's performance and to allocate human resources for actual error correction through manual transcription. With the rapid development of open MLLMs recently, a promising future step is to explore local solutions for fine-tuning and application of the models.

Shintaro Yamada
The University of Tokyo, Japan

Construction and Potential Applications of a Knowledge Graph of Medieval Icelandic Sagas: Aiming Towards Analysis of Problem-Solving Dynamics in Íslendinga Saga

ID: 238 / S01: 5
Doctoral Consortium
Keywords: Icelandic Saga, Knowledge Graph, Problem-Solving, Social Network Analysis

Introduction

In my research, I aim to construct knowledge graphs focusing on the sagas —a collection of medieval Icelandic prose narratives written in Old Norse— and to elucidate the dynamics of problem-solving within these sagas based on that graph. In this presentation, I particularly discuss the construction of the knowledge graph for the saga. The term "saga" in Old Norse refers to "story" or "history." In saga studies, it specifically denotes a collection of prose narratives written in Old Norse. Many of these sagas are considered to have been passed down orally and were transcribed into manuscripts around the 12th to 13th centuries, surviving to the present day. The sagas are categorized into several genres according to their content.

Among them, there is a genre called "contemporary sagas." In most sagas, there is a temporal gap between the period depicted in the saga and the time when it was transcribed into manuscripts. "Contemporary sagas," however, are characterized by the events narrated in the sagas occurring roughly in the same period as when the sagas were written down. For this reason, "contemporary sagas" serve as important materials for understanding the Icelandic society of that time.

Research Objectives

The sagas document countless interactions among people, such as disputes over property, murder, revenge, lawsuits, and marriage. In other words, these interactions depict how problems arose in the daily lives of people at that time, how they dealt with these problems, and what changes occurred in their relationships as a result. By capturing this series of changes as the dynamics of problem-solving and analyzing each interaction process in detail, this study aims to clarify the ideal ways of problem-solving as perceived by the Icelandic people of that time. Regardless of the varying degrees of truth in the sagas' descriptions, contemporary people accepted them as possessing a certain level of authenticity. By unraveling all interactions described in the sagas, this study intends to reveal the ideal forms of problem-solving within them.

This endeavor typifies problem-solving approaches in daily life. By extending this to other narratives and stories from different regions, we can establish a global typology of approaches for problem-solving. Situations requiring problem-solving are not confined within narratives. Learning such knowledge for problem-solving allows us to enjoy it as wisdom opens to us all, transcending time and region.

Materials and Scope

This study focuses on one of the contemporary sagas within the "Sturlunga Saga," specifically "The Saga of the Icelanders." It contains rich documentation of social conflicts and problem-solving instances; thus, it is an ideal case study for analyzing medieval Icelandic dispute resolution patterns. "The Saga of the Icelanders" has two editions of critical texts (published in 1946 and 2021). Based on the 1946 edition, which is already available in digitized files, it consists of 200 chapters. Since the 2021 edition is published only in print, this study first uses the already publicly available digitized critical text (1946 edition). In either edition, the number of characters appearing in the narrative exceeds 100, and many place names appear as well. However, this study also refers to the 2021 edition as appropriate in the research.

Approach

Numerous problem-solving instances occur in the sagas, and manually extracting all of them is time-consuming. Although this research is limited to a single saga, this study has a long-term goal of comparative studies across multiple sagas. Therefore, while maintaining comprehensiveness, it needs to partially streamline the work, for which this study utilizes knowledge graphs.

A knowledge graph allows us to represent relationships between entities as a graph in a machine-readable format. In such graph construction, descriptions based on RDF (Resource Description Framework) are well-known. As mentioned earlier, the sagas depict various human interactions. These interactions have a kind of network structure and can be represented in graph form. Therefore, this study constructs knowledge graphs of the saga's narrative content, keeping the following two points in mind:

1. Continuous Chronological Representation: Represent the events depicted in the sagas continuously while maintaining chronological order.

2. Detailed Expression of Relationships: Express the relationships between characters in detail.

By using query languages such as SPARQL on this knowledge graph, we can extract specific structures and search for complex relational patterns, allowing for comprehensive searches within the graph. Moreover, knowledge graphs are highly extensible; by constructing knowledge graphs of other sagas using a common ontology, they can be integrated. If we can also link images of manuscripts and archaeological materials or any other related data, it can be used as LOD (Linked Open Data) of medieval Icelandic sagas. Therefore, constructing the sagas as a knowledge graph is not only necessary for analysis but will also contribute to building a foundation in saga studies in the long term.

Preliminary Results and Future Prospects

At DHNB2024 last year, I worked on developing a workflow for constructing knowledge graphs of sagas. In constructing the knowledge graph for the sagas, I decided to build two major graphs:

1. Event-Oriented Graph: Representing the content of the text as a graph while maintaining the time information inherent in the sagas' text.

2. Human Relationship-Oriented Graph: Representing the relationships between characters as a graph, referring to genealogical charts and proper noun indexes in the critical texts' appendices.

Among these, I focused on constructing a relatively simpler human relationship-oriented graph. By summarizing human relationships in tabular form using Excel while referring to genealogical charts and proper noun indexes, I demonstrated the procedure to convert this into a knowledge graph.

In my presentation at DHNB2025, based on last year's workflow, I plan to present part of the results from my current work on constructing the event-oriented graph. I employ GPT-4 with Retrieval-Augmented Generation (RAG) to systematically extract named entities, events, and relationships from the saga text. This approach ensures accurate identification of medieval Icelandic names, places, and events from the resource.

This research not only advances the methodology for analyzing the medieval texts but also contributes to the broader field of digital humanities by demonstrating the effectiveness of knowledge graphs in literary studies.

Bibliography

Dodds, Leigh, and Ian Davis. 2022. 'Linked Data Patterns: A Pattern Catalogue for Modelling, Publishing, and Consuming Linked Data'. https://patterns.dataincubator.org/.

Hyvönen, Eero, Petri Leskinen, and Jouni Tuominen. 2023. 'A Data-Driven Approach to Create an Ontology of Parliamentary Work: International Workshop on Semantic Web and Ontology Design for Cultural Heritage'. Edited by Antonis Bikakis, Roberta Ferrario, Stéphane Jean, Béatrice Markhoff, Alessandro Mosca, and Marianna Nicolosi Asmundo. 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), CEUR Workshop Proceedings, November. https://ceur-ws.org/Vol-3540/paper6.pdf.

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.

Lethbridge, Emily. dataARC GitHub repository. (https://github.com/castuofa/dataarc-source).

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. https://doi.org/10.1007/978-3-030-37569-0_2.

Ogawa Jun, et al. 2023 'Collecting Pieces of Historical Knowledge from Documents: Introduction of HIMIKO (Historical Micro Knowledge and Ontology)'. https://doi.org/10.5281/zenodo.8107411.

Yamada Shintaro, Jun Ogawa, Ikki Ohmukai. 2024 'Representing the Íslendinga Saga as Knowledge Graphs of Events and Social Relationships: Developing Workflows Based on a Pilot Case'. https://www.conftool.org/dhnb2024/index.php/YAMADA-Representing_the_%C3%8Dslendinga_Saga_As_Knowledge_Graphs-217.pdf?page=downloadPaper&filename=YAMADA-Representing_the_%C3%8Dslendinga_Saga_As_Knowledge_Graphs-217.pdf&form_id=217&form_version=final.

Aytac Yurukcu
University of Eastern Finland, Finland

Examples form the Translocalis: Cultural Heritage, Narratives, Emotions, Perceptions and Voices of the Finnish Media, People, and Soldiers on the Imperial War

ID: 307 / WS04A: 2
Explorations of the dynamics of cultural phenomena in text corpora
Keywords: Emotions, war experiences, newspapers, local letters, ideas, digital humanities.

The nineteenth century was a challenging period not only for the Ottoman Empire, which Russia condescendingly referred to as "The Sick Man of Europe," but also for Russia, which faced all the major powers in the Crimean War. These empires fought nine times between the beginning of the seventeenth century and the Russo-Turkish War of 1877–78. The war had far-reaching consequences in the Balkans and Caucasus and was unique in the way journalists portrayed it for Europeans, Russians, Turks, and Balkan nations. WarThe media, journalists, military attaches, and correspondents from various nations shaped war news, attracting a large audience and often deeply engaging people both emotionally and intellectually. However, the conflict also had a significant impact on the ethnic majorities and peripheral minorities of the Russian Empire, including Finns, Estonians, and Poles, who served in the armThis research will examine over 60 published letters from readers and soldiers in 1877-78, sourced from the Translocalis Database [1]. UsiReaders and soldiers from the war zone sent local letters to newspapers as source material, providing details on the perspectives of society, people, and war, and sharing the community's experiences with the war. study posits a significant hypothesis that Finland's emerging notion of a distinct state and nationhood was influenced by the wartime events, as evidenced by particular instances from readers' letters, news coverage of the war in newspapers, and tales of troops. The qualitative content analysis and digital humanities tools (https://korp.csc.fi/korp/, AntConc tool, and digitalised newspaper collection of Finnish National Archives https://digi.kansalliskirjasto.fi/collections?id=742) will be implemented to analyse the written texts by key themes: society, soldiers, solidarity, narratives, war news, and enemy images.

[1] Translocal Database developed by the Academy of Finland Centre of Excellence in the History of Experiences (HEX), https://digi.kansalliskirjasto.fi/sanomalehti/binding/431835?page=3 is an online repository for reader correspondence sent in various places and published in Finnish newspapers from 1844 till 1885, it comprises 71,826 reader letters originating from Finland and several different countries. Translocalis data is a sample of early part of the formation of Finnish civil society. In a way it is the social media of its time. It is internationally unique source of data and especially Finnish phenomenon.

Aytac Yurukcu
University of Eastern Finland, Finland

Estonian and Finnish Soldiers’ War Song’s during the Imperial War in the Balkans in 1877-1878

ID: 308 / WS10: 9
Tradition Archives Meet Digital Humanities II
Keywords: War, Songs, Soldiers, War Experiences, Letters.

The Russian and Turkish empires fought nine times between the beginning of the 17th century and the Russo-Ottoman War of 1877–78. Following the 1875–1876 wars between Serbs and Turks, this war of 1877–78 had far-reaching consequences in the Balkans and Caucasus for Russians, Turks, and Balkan nations. But the war also had a significant impact on the ethnic majorities and peripheral minorities of the Russian Empire, such as Finns, Estonians, Latvians, Lithuanians, and Prussians. This paper aims to analyze the war songs of the Finnish and Estonian soldiers who participated in the Russian-Ottoman War in the Balkans in 1877–1878. There were 1114 Finnish and 2636 Estonian soldiers in different locations during the war. This study utilized the old war songbooks belonging to both Finnish and Estonian troops. Particularly with one Estonian and songbook published in 1877 (a) and one Finnish songbook published in 1878 (b), (Wene ja Türgi Sõa Laulud nastak 1877: „Meie mees, fui sõtta lääb, Siis ta aun ja wõitu saab." [Russian and Turkish Folk Songs from 1877: „Our man, if he goes to war, will then gain honour and victory"], and Kaunis Laulu, Turkin ja Wenäjän sodasta [Kaunis Laulu, Turkin ja Wenäjän sodasta],) from the letters of soldiers and their shared experiences about the imperial and religious war between Christians and Muslims (Cross and Crescent) by using translation and text analyses methods. By implementing the Orange and AntConc tools, we can use text analyses and mining methods to comparatively assess the folk cultures of Finland and Estonia during the war. This comparative assessment will not only highlight the distinct cultural narratives that emerged in response to the conflict but also reveal the underlying themes of resilience and unity among the soldiers. Additionally, the insights gained from this analysis can contribute to a broader understanding of how wars shape cultural identities and influence the collective memory of nations.

Bibliography

F.B. Wene ja Türgi Sõa Laulud nastak 1877: „Meie mees, fui sõtta lääb, Siis ta aun ja wõitu saab." [Russian and Turkish Folk Songs from 1877: „Our man, if he goes to war, will then gain honor and victory"], Printed, H. Gressel's letters, Tallinnas, 1877. (With ten folk war songs published from the letters of the soldiers).

U. Leppänen, *Kaunis Laulu, Turkin ja Wenäjän sodasta [Kaunis Laulu, Turkin ja Wenäjän sodasta],* Oulusta, from the Printing House of Joh. Bergbahl, 1878.