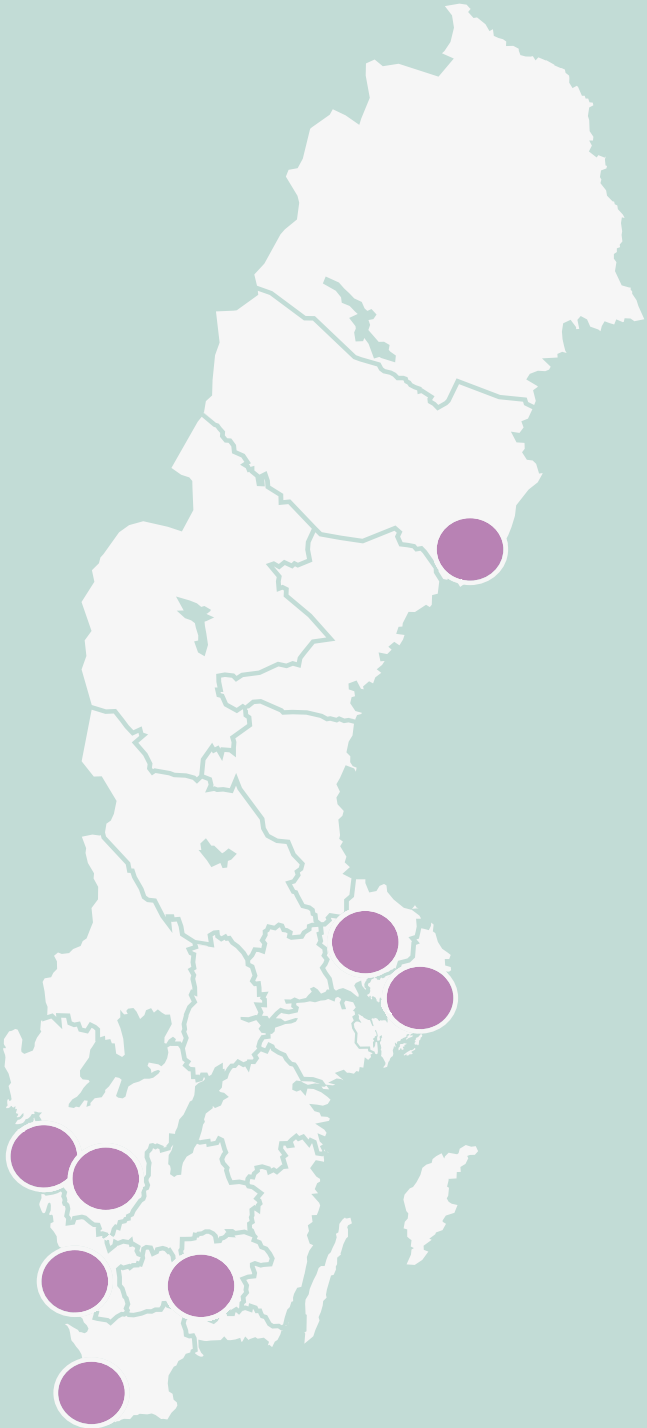
HiC 2025

HumInfra Conference

12-13 November, 2025

Stockholm, Sweden



HUM INFRA

Swedish Research Council

Proceedings of the 2nd Huminfra Conference

HiC 2025

Editors
Harko Verhagen, Mats Fridlund, Magnus Nermo,
Frantzeska Papadopoulou Skarp, Susanne Tienken,
Andreas Widholm, and Anna Blåder

November 12-13, 2025
Stockholm, Sweden

Published by

NEALT Proceedings Series 60
D-Space at Tartu University Library
ISSN: 1736-8197
eISSN: 1736- 6305
ISBN: 978-99-0853- 677-4

Preface

On November, 12-13, 2025, the Stockholm University unit for digital human science[1] organized the second Huminfra[2] conference with the aim to showcase the variety of infrastructural tools, resources and initiatives aimed at supporting digital and experimental research in the Humanities as well the Social Sciences and Law (together constituting digital human science (DHS)). Topics of interest included amongst others descriptions and showcasing of digital archives & repositories and experimental tasks, description of DHS centres, networks, and related projects, DHS expertise and expertise in experimental human science, practical aspects of use of DHS tools/services/resources, user training and education (e.g., "how-to" tutorials), dissemination of digital and experimental infrastructures, data, software, and/or research output in and beyond knowledge institutions, and emerging trends and future directions in DHS infrastructure development (e.g., AI).

Submissions were solicited from within the HumInfra network, associated researchers and other colleagues as well as the Dariah[3] network. Submissions were either abstracts, mainly aiming for poster presentations or tool demonstrations, and short papers, aiming for oral presentations.

We received 54 submissions in total, mainly from Sweden but also from Finland, Iran, and the Netherlands. After a review process in which all submissions received at least 2 reviews, based on which the organization committee took its decisions, we accepted 15 short papers (up to 6 pages) and 25 abstracts for presentation during the conference (75% acceptance rate). The accepted submissions concerned research resources such as hardware and software (or combinations thereof), datasets, but also for instance ideas on how research engineer career paths could be envisioned and supported alongside the regular academic career paths. The conference organized these in sessions based on research domain, while the proceedings are ordered in alphabetical order of the last name(s) of the first author. The present volume collects 14 accepted short papers. Abstracts are collected in a separate non-archival volume available at https://www.huminfra.se/HiC-2025.

Apart from the submissions, we also had 2 invited keynote speakers at HiC2025:

- Vicky Garnett

  Research Fellow in the Trinity Centre for Digital Humanities at Trinity College Dublin, and holds a PhD in Sociolinguistics with a strong interest in research ethics, and is the Training and Education Officer for DARIAH-EU. In 2024, DARIAH celebrated its 10 year anniversary as an ERIC (European Research Infrastructure Consortium). In 2025, DARIAH turned to the future as we began drafting our next Strategic Plan (2026-2030). This talk will reflect on some of the issues that have arisen since 2014 specific to training and education within the European landscape, and how the DARIAH community has both informed and reacted to those issues. We will look at the lessons learned along the way, and what opportunities await national consortia to engage with shaping how European Research Infrastructures such as DARIAH meet the future needs of their communities.

- Åsa M Larsson

  Head of Unit for Technology and Digital Mediation at the Swedish National Heritage Board. PhD in Archaeology, now working with developing digital archaeological processes and implementing FAIR data principles within the cultural heritage sector and research in general. On the Operational Board for Swedigarch Research Infrastructure for Digital Archaeology as Leader of Module 2: FAIR and Linked Data. The Heritage Board is developing K-Samsök, the national aggregator for heritage data, allowing for more advanced ways of indexing, linking, and publishing data from various databases at museums, archives, universities, and government

[1] https://su.se/dhv

[2] Huminfra (htttps://huminfra.se) is a Swedish national research infrastructure supporting digital and experimental research in the Humanities by providing users with a single entry point for finding existing Swedish materials and research tools, as well as developing national methods courses.

[3] https://www.dariah.eu/

agencies. The new aggregator will be based on the CIDOC-CRM data model and compatible with general standards for data, to ensure compatibility with international infrastructures. It is also implemented on a graph database to meet the more advanced needs among today's researchers in Digital Humanities and Data Sciences.

HiC2025 and Huminfra gratefully acknowledge funding and support from the Swedish Research Council (grant number 2021-00176) and all the partner institutions: Lund University, Umeå University, University of Gothenburg, KTH Royal Institute of Technology, the National Library of Sweden, Stockholm University, the Swedish National Archives, Uppsala University, the Swedish School of Library and Information Science, Linnaeus University, and Halmstad University. Special thanks to the organization committee for the support in practical matters as well as putting the programme together, and the programme committee for their dedication and efforts in the reviewing of the submissions.

Stockholm, December 7, 2025

Harko Verhagen
General Chair of HiC2025

Program Committee

David Alfter, University of Gothenburg
Marie Dubremetz, Uppsala University
Sara Ellis Nilsson, Linnaeus University
Mats Fridlund, University of Gothenburg
Koraljka Golub, Linnaeus University
Ashely Green, University of Gothenburg
Marianne Gullberg, Lund University
Fredrik Hanell, Linnaeus University
Olof Karsvall, The Swedish National Archives
Julia Kuhlin, Linnnaeus University
Matti La Mela, Uppsala University
Evelina Liliequist, Humlab, Umeå University
Gustaf Nelhans, University of Borås
Diederick C. Niehorster, Lund University
Tomas Nilson Halmstad university
Frantzeska Papadopoulou Skarp, Stockholm University
Justyna Sikora, National Library of Sweden
Daniel Sundberg, Linnaeus University
Susanne Tienken, Stockholm University
Matteo Tomasini, University of Gothenburg
Harko Verhagen, Stockholm University
Elena Volodina, University of Gothenburg
Rebecka Weegar, Umeå universitet
Andreas Widholm, Stockholm University

Organizing Committee

Harko Verhagen, Department of Computer and Systems Sciences & Digital Humanvetenskap unit (DHV), Stockholm University (General Chair)
Mats Fridlund, Gothenburg Research Infrastructure in Digital Humanities (GRIDH), University of Gothenburg
Magnus Nermo, Department of Sociology, Stockholm University
Frantzeska Papadopoulou Skarp, Department of Law & DHV, Stockholm University
Susanne Tienken, Department of Slavic and Baltic Studies Finnish Dutch and German & DHV, Stockholm University
Andreas Widholm, Department of Media Studies & DHV, Stockholm University
Anna Blåder, Lund University Humanities Lab (Humanistlaboratoriet), Lund University

Contents

A Distant Technology? Experiments with a Generative Model for Retouching Noisy Newspaper OCR

Daniel Brodén[1], Lina Samuelsson[2], David Alfter[1], Johan Malmstedt[1,3]

[1] *University of Gothenburg, Box 200, Gothenburg, 405 30, Sweden*
[2] *Mälardalen University, Box 325, 631 06 Eskilstuna, Sweden*
[3] *Harvard University, MA 02138, Cambridge, USA*

Abstract

This paper explores the use of generative language models to enhance digitized historical newspaper text. While large language models offer new means of addressing noisy OCR, their opaque, probabilistic processes raise epistemological concerns. Within the project *The Order of Criticism Revisited*, which integrates literary and computational approaches to Swedish criticism, we tested GPT-4o to "retouch" OCR data from the National Library of Sweden using zero-shot prompting. Comparisons with flawed OCR outputs and manually transcribed texts show that the model produced more legible versions, often closer to the originals than the raw OCR. This indicates potential for improving the quality of digitized sources and enabling more robust large-scale analysis. At the same time, drawing on the notions of artificial communication and distant technology, we argue that such models extend analytical capacity while creating perceptual and methodological distance. Their outputs, better seen as probabilistic "retouching" than correction or reconstruction, weaken the indexical link to original sources.

Keywords

Generative models, OCR, digital epistemology

1. Introduction

Large language models and generative models, like GPT [1] are currently reshaping how researchers interact with source materials [2]. Yet, because these models rely on opaque computational processes, it is difficult to evaluate the reliability and validity of their outputs [3]. At the same time, within digital humanities, experimentation as a scholarly principle encourages reflective inquiry into the potential of digital tools and methods [4]. In this spirit, rather than seeking systematic methodological assessment, this paper reflects on both the capacity and the implications of using a generative model to enhance the data quality of familiar source material more broadly.

The research project *The Order of Criticism Revisited* (2020–2025) [5] explores how "traditional" literary scholarship intersects with computational methods, building on materials and results from an earlier study of literary criticism in Swedish press [6]. Within the project, we collected and annotated approximately 5,800 book reviews, primarily from the National Library of Sweden's (*Kungliga biblioteket*, hereafter KB) newspaper collections but digitization shortcomings, including optical character recognition (OCR), complicate their usefulness [7] [8] [9] These limitations highlight both the need for careful curation for research [10] and automated OCR mitigation of structural flaws [11].

Our inquiry is framed by a two-part question: To what extent can a generative language model make noisy OCR in digitized Swedish newspapers more usable, and what uncertainties arise in doing so? Specifically, we explore using GPT-4o for "generic" OCR cleaning through zero-shot prompting, directing the model to perform the task based on patterns inferred from its training data rather than through the more resource-intensive process of fine-tuning on a specific dataset. We chose GPT-4o as our primary model since, at the time of writing, it was among the most advanced and widely adopted, showing particular advantages in our tests (for comparison, we also used Claude 3.7 Sonnet but

Huminfra Conference 2025, Stockholm, 12-13 November 2025.

daniel.broden@lir.gu.se (D. Brodén); lina.samuelsson@mdu.se (L. Samuelsson); david.alfter@lir.gu.se (D. Alfter); johan.malmstedt@lir.gu.se (J. Malmstedt)

encountered limitations, as it failed to return complete outputs due to input-length restrictions). Given both the proprietary status of GPT-4o and copyright constraints of KB's collections, we limited our focus to reviews published in 1906.

To frame our inquiry, we first draw on the theoretical notions of artificial communication and distant technology to highlight how generative models, by virtue of their opaque and probabilistic nature, add a further layer of technological distance from research materials. We then discuss our experiment of applying GPT-4o to "retouch", rather than reconstruct, low-quality OCR in KB's collections. This design allows us to reflect on both the model's capacity to enhance text quality and the epistemological uncertainties that arise when generated outputs further weaken the indexical link to original sources. The paper concludes that although the model can substantially improve degraded OCR texts, its use simultaneously introduces a methodological distance that demands thorough reflection.

2. Digital Epistemology

Digital epistemology scholar Jonas Ingvarsson contends that digitization challenges the foundations of academic practice "not only by the appearance of new tools and objects, but by the fact that our modes of thought and our way to structure data and knowledge are changing" [12]. From this perspective, digital forms of expression provide scholars with different modes of engaging with research materials.

While data-intensive studies of literary criticism and newspaper collections tend to emphasize the analytical potential of computational methods rather than their epistemological implications [13] [14] [15], we have argued that text-data visualizations of previously analyzed book reviews can create a defamiliarizing effect, opening new perspectives about the material [16]. With generative models, however, a different kind of distancing effect emerges. These models involve forms of methodological uncertainty that partly differ from those encountered in the visualization techniques we earlier employed, such as TF-IDF analysis [16]. While GPT-4o generates text that resembles human writing and conveys a sense of familiarity with the material, its outputs remain products of probabilistic processes that scholars have described as largely opaque [17] [18].

2.1. Artificial Communication

At first glance, GPT-4o might seem to lessen the distance between researcher and data, thanks to its intuitive interface and human-like output, especially when compared with more conventional natural language processing (NLP) and "AI" techniques. Yet, as sociologist Elena Esposito points out: "what we can observe in interactions with [these] algorithms is not necessarily an artificial form of intelligence, but rather an artificial form of communication. Intelligence and communicative capacity are not the same thing" [19]. Esposito further argues that "algorithms learn not to think but to participate in communication, that is, to (artificially) develop an autonomous perspective that allows them to react appropriately and generate information in their interaction with other participants" [19].

Because generative models simulate the exchange and interpretation of information within linguistic distributions shaped by training data, their outputs should not be conflated with intelligence-based data processing [20]. In this, generative models partly create a different sort of distance from research results than traditional NLP methods. The latter may seem opaque to those without technical expertise, but they remain accessible and, to some extent, interpretable for those with the necessary knowledge. Generative models, such as GPT-4o, by contrast, mediate methodological distance for virtually all. While these models may display impressive capacity across a variety of research tasks, the computational processes that drive their responses are largely unreachable, even if new advances in interpretability and reverse engineering are beginning to provide insights into their internal workings [21].

2.2 Technological Distancing

On some level, this touches on psychologist and philosopher of technology Robert Romanyshyn's idea that modern technologies not only expand human capabilities but also reshape perception by distancing us from the objects of our attention [22]. Although Romanyshyn's account is primarily

concerned with the ways in which modern technology generates human detachment from the world (a point that remains debated), it more broadly highlights the ways in which technologies can appear to reduce our sense of limitation while simultaneously altering how we engage with the world [23]. In our context, this idea of distant technology seems relevant for understanding the relationship between researcher and research material, which becomes mediated through the use of a generative model.

The concept of distant technology thus provides a basis for reflecting on the dual methodological character of generative models. Within scholarly practice, models such as GPT-4o can function as powerful tools for a wide range of tasks, yet they also obscure the processes by which their output is produced. In this way, these models embody a form of technologically mediated engagement that both expands scholarly capacity and simultaneously introduces an additional layer of distance between researcher and material. While other OCR methods are likewise probabilistic and often black boxed, GPT-4o's smooth, high-quality, and seemingly error-free output arguably heightens the issue. Taken together, these theoretical perspectives highlight both the ambivalent character of the model as a research instrument and the need for thorough reflection.

3. Testing GPT-4o on Newspaper Text

To demonstrate the scope of the OCR-related problems with KB's collections, we can highlight the following excerpt from a review in the newspaper *Arbetet* of Gustaf af Geijerstam's novel *Farliga makter* [*Dangerous Powers*] (1906), which we have manually transcribed from the original print edition:

> Geijerstam eger på en gång vetenskapsmannens lugna objektivitet och diktarens förmåga att ge lif åt sina gestalter, och härmed sammanhänger den manligt okonstlade, man vore frestad säga sakliga stil, som präglar mycket af hans produktion och i all synnerhet ”Farliga makter”. Han dekorerar icke sina figurer med förträfflighetens epiteter, lika litet som han förlöjligar dem; han vet att det enda, hvaraf man kan sluta till en människas karaktär är hennes handlingar, och utan att försmå den psykologiska analysen låter han dock sina figurer först och sist uppträda som handlande individer [24].

This can be compared with the error-riddled passage from KB's OCR-processed version:

> Geijerstam eger på en gång vetenskapsmannens lugna ob jlek livi te t och diktarens förmåga att ge lif åt sina gestalter, och härmed sammanhänger denmanligt okonstlade, man vore frestadsäga sakliga stil, som käglar mycketaf hans produktion och i all synnerijftt -sdjaxli cj^.maktfcDiwi 4’ * »Ia’icke sina figurer med förträfflighetensepiteter, lika litet som han förlöjligardem; han vet att det enda, hvarafman kan sluta till en människas karaktär är hennes handlingar, och utan attförsmå den psykologiska analysen låter han dock sina figurer först ochsist uppträda som handlande individer.

3.1. Broken Indexicality?

While OCR errors may or may not impede downstream computational tasks, depending on the degree to which those tasks rely on orthographic precision, prior studies have demonstrated the detrimental impact of source text quality on, for instance, topic modeling. Nevertheless, whereas the analytical consequences of noisy OCR have conventionally been mitigated through manual curation, addressing such issues by means of a generative model entails a form of textual fabrication. One could debate which term most accurately describes this process, but we have tentatively chosen “retouching” as a more precise alternative, since it suggests both reuse and modification. Labels such as “correction” [26], “reconstruction”, or “restoration” seem a bit misleading, as the generated text is not derived from the original printed source but from a flawed digitized variant.

In effect, no clear indexical relationship with the source material remains [27]. Notably, OCR already performs a reduction of the document's semiotic complexity by disregarding qualities like letter shape and whitespace and thus transforming from image to text [28]. Generative retouching extends this abstraction one step further. Rather than repairing the losses introduced by OCR, it takes the already filtered text as its input for complex vectorial calculations that operate probabilistically on the level of meaning instead of on the level of the singular letter, effectively layering another transformation on top of the initial transcription. The unprocessed OCR output, despite its flaws, still retains an indexical resonance with the page as an artifact, since its errors often betray the contours of the original print. The retouched variant, by contrast, suppresses even those traces. What remains is not a causal imprint but a statistically inferred approximation, where referentiality is mediated twice over.

3.2 Significant or Negligible Discrepancies?

Notably, GPT-4o produces these texts not by recognising letterforms from pixel values or singular glyph information, but by mobilizing large-scale symbolic relations acquired during training. Instead of reconstructing the original texts through direct reference to the newspaper artifacts, it generates versions that are linguistically and contextually coherent, drawing on patterns of usage across vast corpora. In this sense, the model does not recreate the original material but provides an approximation mediated by the statistical and symbolic structures of language itself. Consequently, one cannot expect the generated texts to correspond exactly to the originals. Still, when we turn to the passage cited above, the retouched version produced by GPT-4o reads as follows:

> Geijerstam *äger* på en gång vetenskapsmannens lugna objektivitet och diktarens förmåga att ge *liv* åt sina gestalter, och härmed sammanhänger den manligt okonstlade, man vore frestad säga sakliga stil, som präglat mycket av hans produktion och i all synnerhet *fallet där makterna är mäktigast*. Han *prydar* icke sina figurer med förträfflighetens *epitet*, lika litet som han förlöjligar dem; han vet att det enda, *varav* man kan sluta till en människas karaktär är hennes handlingar, och utan att försmå den psykologiska analysen låter han dock sina figurer först och sist uppträda som handlande individer. [italics indicate divergence between the GPT-4o output and the printed originals]

Importantly, the generated text does not match the original exactly. There are discrepancies in spelling (GPT's version is modernized), a plural form has been rendered as singular and for illegible portions of the text the model has hypothesized that the title of the novel *Farliga makter* [Dangerous Powers] should instead be rendered as "fallet där makterna är mäktigast" ("the case where the powers are at their strongest"). While some differences relate only to spelling, others are more substantial and affect the process of meaning-making itself. A sentence-by-sentence comparison of the smaller set of retouched reviews with the original texts also certifies that this is the kind of typical differences we can see in the material. Since the output neither reproduces the original text with complete accuracy nor preserves a record of its alterations, it would be risky to base analytical interpretations solely on the retouched versions. At the same time, however, the GPT-4o output often appears closer to the original than the OCR-processed text found in KB's digitized newspaper collections.

3.3 Transparency and Uncertainty

A manual evaluation comparing the transcribed reviews with both KB's OCR versions and the GPT-4o outputs shows that the latter align much more closely with the transcriptions. To assess the OCR retouching process, we analyzed both the full corpus (n=394) and a manually transcribed subset (n=21). This involved calculating the Levenshtein distances between the noisy OCR text and its modified version, as well as between both versions and the manual transcriptions. The Levenshtein distance measures the number of operations (insertion, deletion, or substitution of characters) required to transform one string into another, providing a measure of textual change. A low score indicates minimal alterations, whereas a high score signals greater divergence. On average, the OCR retouching required 320 operations per text (95% CI: 253-386, n=394). In the manually transcribed subset, the retouched

texts were, on average, closer to the manual transcriptions than the original OCR versions, with the noisy OCR texts approximately 104 operations further from the gold standard (95% CI: 64-143, n=21).

Nevertheless, this raises epistemological concerns, as we are dealing with texts that appear more accurate and usable for computational analysis yet are not derived from, and remain detached from, the original print. In this sense, GPT-4o's contribution to OCR cleaning does more than enhance access to historical sources; it also introduces a layer of epistemological uncertainty. In the light of Esposito's argument that generative models are better understood as forms of artificial communication rather than intelligence, this retouching can be seen as a simulation that reshapes noisy OCR output while obscuring the processes behind it. At first glance, the result seems to bring us closer to the historical text, yet it introduces a perceptual and methodological distance. The improved legibility comes at the expense of traceability: without direct comparison to the print sources, similar to other OCR-correction processes, it is impossible to assess how closely the model's output reflects the originals. Rather than restoring the past, GPT-4o produces a plausible representation of it, underscoring the need for deeper scrutiny.

4. Conclusions

In this paper, we have experimented with and reflected on the application of a generative language model to Swedish literary criticism from 1905–1906. The experiment demonstrated the model's capacity to improve OCR-degraded texts, providing a stronger basis for data-driven analysis. At the same time, we showed that this process weakens the connection to the original sources and introduces epistemological uncertainty. Although the generated texts often resemble the originals more closely than the noisy OCR versions, they are best regarded as probabilistic retouching, rather than restoration, since they lack an indexical link to the original material and can no longer reliably point back to it.

Building on the idea that AI functions as a communicative system producing plausible outputs without semantic understanding, based on probabilistic pattern recognition, we have emphasized that a generative model like GPT-4o may perform what resembles high-quality OCR-cleaning while remaining methodologically distant. This distance arises primarily from the opacity of its computational processes, which complicates methodological transparency and calls for reflection on the interpretive gaps such systems introduce. As generative models become more prominent in humanities research, the task is not only to refine their practical applications but also to develop strategies for mitigating methodological distance and a theoretical awareness of the conditions under which these tools operate.

Acknowledgements

This paper was prepared as part of the research project *The Order of Criticism Revisited* (2020–2025), funded by Riksbankens Jubileumsfond (RJ), grant no. MXM19-1096:1.

References

[1] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al. "Language Models are Few-shot Learners." Advances in Neural Information Processing Systems 33 (2020): 1877–1901.

[2] C.-E. González-Gallardo, T.T.H. Hahn, A. Hamdi, A. Doucet, "Leveraging Open Large Language Models for Historical Named Entity Recognition", in: A. Antonacopoulos, A. Hinze, B. Piowowarski, M. Coustaty, G.M. Di Nunzio, F. Gelati, N. Vanderschantz (Eds.), Linking Theory and Practice of Digital Libraries, TPDL 2024: 28th International Conference on Theory and Practice of Digital Libraries, Ljubljana, Slovenia, September 24–27, 2024, Proceedings, Part I, Springer, 2024.

[3] R. Bommasani, D. A. Hudson, E. Adeli, R. Altman, S. Arora, S. von Arx, M. S. Bernstein, J. Bohg, A. Bisselut, E. Burnskill, et al. 2021. "On the Opportunities and Risks of Foundation Models", Preprint, arXiv (2021). URL: https://arxiv.org/abs/2108.07258.

[4] J. Drucker, SpecLab: Digital Aesthetics and Projects in Speculative Computing, University of Chicago Press, Chicago, 2009.

[5] J. Ingvarsson, D. Brodén, L. Samuelsson, V. Wåhlstrand Skärström, N. Zechner, “The New Order of Criticism: Explorations of Book Reviews Between the Interpretative and Algorithmic”, in: K. Berglund, M. La Mela, I. Zwart (Eds.), The 6th Digital Humanities in the Nordic and Baltic Countries Conference (DHNB 2022), CEUR-WS.org (2022), pp. 228–234.
[6] L. Samuelsson, Kritikens ordning: Svenska bokrecensioner 1906, 1956, 2006, Bild, text & form, Karlstad, 2013.
[7] D. Brodén, L. Samuelsson, N. Zechner, J. Ingvarsson, A. Karimi, “Between the Arduous and the Automatic: A Comparative Approach to the Challenge of Identifying Book Reviews in Swedish Newspapers”, in: O. Holownia, E.S. Sigurðarson (Eds.), Digital Humanities in the Nordic and Baltic Countries 2024, May 27–31, 2024, Reykjavik, Iceland, University of Oslo Library (2025), pp. 1–12.
[8] J. Jarlbrink, P. Snickars, C. Colliander. “Maskinläsning: Om massdigitalisering, digitala metoder och svensk dagspress.” Nordicom Information 38.3 (2016): 27–40.
[9] L. Börjesson, C. Haffenden, M. Malmsten, F. Klingwall, E. Rende, R. Kurtz, F. Rekathati, H. Hägglöf, J. Sikora. “Transfiguring the Library as Digital Research Infrastructure: Making KBLab at the National Library of Sweden.” College & Research Libraries 85.4 (2024): 564–582.
[10] J. Sikora, C. Haffenden, “AI, Data Curation and the Readiness of Heritage Data: Exploring the Swedish Newspaper Archive at KBLab”, in: E. Volodina, G. Bouma, M. Forsberg, D. Kokkinakis, D. Alfter, M. Fridlund, C. Horn, L. Ahrenberg, A. Blåder (Eds.), Proceedings of the Huminfra Conference (HiC 2024), Linköping Electronic Conference Proceedings (2024), pp. 60–66.
[11] V. Löfgren, D. Dannélls, “Post-OCR Correction of Digitized Swedish Newspapers with ByT5, in: J. Bizzoni, S. Degeatano-Ortlieb”, in: A. Kazantseva, S. Szpakowicz (Eds.), Proceedings of the 8th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature (LaTeCH-CLfL 2024), Association for Computational Linguistics (2024), pp. 237–242.
[12] J. Ingvarsson, Towards a Digital Epistemology: Aesthetics and Modes of Thought in Early Modernity and the Present Age, Palgrave Macmillan, Cham, 2020.
[13] T. Underwood, Ted, Distant Horizons: Digital Evidence and Literary Change, University of Chicago Press, Chicago, 2019.
[14] A. Piper, Can We Be Wrong? The Problem of Textual Evidence in a Time of Data, Cambridge University Press, Cambridge, Mass, 2020.
[15] J. Brottrager, A. Stahl, A. Arslan, U. Brandes, T. Weitin. “Modeling and Predicting Literary Reception. A Data-rich Approach to Literary Historical Reception.” Journal of Computational Literary Studies 1.1 (2022): 1–27.
[16] D. Brodén, J. Ingvarsson, L. Samuelsson, V. Wåhlstrand Skärström. “Visualization as Defamiliarization: Mixed-methods Approaches to Historical Book Reviews.” Journal of Computational Literary Studies 3.1 (2024): 1–26.
[17] J.M. Mathews, 2022. “Some Critical and Ethical Perspectives on the Empirical Turn of AI Interpretability.” Technological Forecasting and Social Change 174: 121209.
[18] J. Hewitt, R. Geirhos, B. Kim, “We Can’t Understand AI Using Our Existing Vocabulary”, Preprint, arXiv (2025). URL: https://arxiv.org/abs/2502.07586.
[19] E. Esposito, Artificial Communication: How Algorithms Produce Social Intelligence, MIT Press, Boston, 2022.
[20] M. B. Fazi. “Can a Machine Think (Anything New)? Automation Beyond Simulation.” AI & Society: Knowledge, Culture and Communication 34.4 (2019): 813–24.
[21] A. Galgoon, K. Filom, A.R. Kannan, Mechanistic Interpretability of Large Language Models with Applications to the Financial Services Industry, Preprint, arXiv (2024). URL: https://arxiv.org/abs/2407.1121
[22] R. Romanyschyn, *Technology as Symptom and Dream*, Routledge, London and New York, 1989.
[23] J.P Telotte, A Distant Technology: Science Fiction Film and the Machine Age, Wesleyan University Press, Middletown, 1999.
[24] B. L. [Bengt Lidforss], “Nya böcker”, Arbetet (1906-03-10).
[25] https://tidningar.kb.se/s3n6n0hdqhjb8g57/part/1/page/2?q=geijerstam%20AND%20%22farliga%20makter%22%20AND%20%22tre%20m%C3%A5nader%20ha%22

[26] E. Boros, M. Ehrmann, M. Romanello, S. Najem-Meyer, F. Kaplan. “Post-Correction of Historical Text Transcripts with Large Language Models: An Exploratory Study”, in: *Proceedings of the 8th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature (LaTeCH-CLfL 2024* (2024), pp. 133–59.
[27] J. Sterne. “Analog”, in: B. Peters (Ed.) Digital Keywords: A Vocabulary of Information Society and Culture, Princeton University Press, New Jersey, 2016, pp. 31–44.
[28] M. P. Eve, Theses on the Metaphors of Digital-textual History. Stanford University Press, Stanford, CA, 2024.

Appendix: Prompts for OCR retouching

System Prompt (original in Swedish)

Du är en expert inom OCR-korrigering av äldre svenska texter, med särskilt fokus på dokument från omkring år 1906. Din uppgift är att noggrant identifiera och korrigera OCR-fel i en text samtidigt som du bevarar originalets stil, historiska språkbruk och mening. När du utför korrigeringarna ska du:

Återge den ursprungliga tonen och språkliga karaktären som var typisk för tiden, inklusive äldre stavnings- och grammatikformer.

Rätta uppenbara fel i teckenigenkänning, såsom felaktiga bokstäver, orddelar eller interpunktion, utan att modernisera språket.

Säkerställa att korrigeringarna förbättrar läsbarheten och den språkliga korrektheten, samtidigt som den historiska känslan bevaras.

Om du är osäker på en korrigering, markera felet eller lämna det oförändrat med en anteckning för vidare granskning.

Arbeta metodiskt och noggrant, med beaktande av den kontext och stil som är typisk för svensk skrift från början av 1900-talet.

Använd denna vägledning för att producera en korrigerad version av OCR-texten med hög noggrannhet och respekt för den historiska textens ursprungliga uttryck.

Returnera endast den korrigerade texten, utan några kommentarer eller ytterligare förklaringar.

User Prompt (original in English)

OCR-correct this text. Do not shorten the text:

A Frame-Semantic Parsing Plugin for Swedish Research Infrastructure*

Dana Dannélls[1,*,†], Shafqat Virk[†]

Språkbanken Text, University of Gothenburg, 405 30 Gothenburg, Sweden

Abstract

We present the development of a frame-semantic parsing plugin for Sparv – an annotation pipeline for Swedish. The plugin integrates a frame-semantic parser into the annotation pipeline, enabling the automatic identification of frames evoked by lexical units and the assignment of frame elements to their corresponding arguments in text. Designed to operate seamlessly within the infrastructure, the plugin takes raw text as input, and outputs semantic role information in a standardized format compatible with other annotation layers. This implementation demonstrates how frame-semantic analysis can be made available as an additional corpus annotation layer, enriching texts with structured semantic representations that go beyond syntactic or lexical features. By providing access to semantic role information, the plugin can support a wide range of research applications, including semantic search, discourse analysis, and investigation of meaning variation in language use.

Keywords

Digital humanities, Frame semantic parsing, Swedish infrastructure

1. Introduction

Språkbanken Text (SBX), the Swedish research infrastructure for language technology, develops freely available digital research platforms. One of these platforms called Strix [1] is a document-centric platform designed for researchers in the humanities, enabling them to create, annotate, and analyze documents in meaningful and semantically informed ways. The platform, along with other SBX platforms, is built on the annotation pipeline, Sparv [2], which makes it possible to analyze digital material in depth and to formulate nuanced research questions about the data. For example, semantic analysis is performed with the help of lexical-semantic databases such as Saldo [3] and Swedish FrameNet [4]. This process produces semantic representations that make it possible to identify words according to their senses within a conceptual network. As a results, it is possible to highlight words according to the semantics of LEADERSHIP (Figure 1a) or display the frequencies of three major cities in Sweden mentioned in the same collection over time (Figure 1b), or visualize all the geographical locations referenced in the collection (Figure 1c).

Currently, the semantic analysis operates at the word level, that is, it relies on a simple word-matching approach against lexical resources, and does not provide deeper sentence-level interpretation. As a result, questions about which titles or roles are mentioned in specific places and periods, and how these occurrences might relate to historical developments, cannot be systematically explored. To move beyond word-level analysis and capture the underlying meaning within sentences, semantic role labeling (SRL) – the task of automatically assigning semantic roles – can be applied. Recent advances in transformer-based language models have made it possible to train an SRL model for Swedish [5]. In this paper, we describe how this model was integrated into Sparv to enhance semantic analysis. We highlight both the technical implementation and the potential applications for digital humanities research, demonstrating how SRL

Huminfra Conference 2025, Stockholm, 12–13 November 2025.

*You can use this document as the template for preparing your publication.

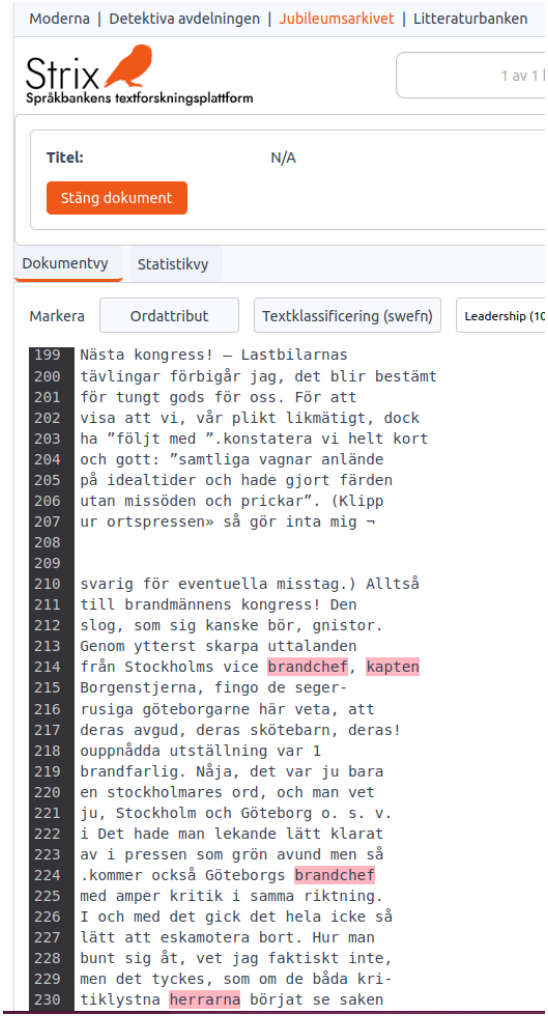
*Corresponding author.

†These authors contributed equally.

dana.dannells@svenska.gu.se (D. Dannélls); shafqat.virk@svenska.gu.se (S. Virk)

0000-0002-3338-2979 (D. Dannélls); 0000-0002-5030-9191 (S. Virk)

integration can facilitate new forms of text analysis, improve interpretability, and enable more nuanced exploration of linguistic and cultural phenomena.



(a) Analysis according to the LEADERSHIP frame in Swedish FrameNet

(b) Frequencies of Stockholm, Gothenburg, and Malmö in the collection over time

(c) Geographic locations mentioned in the collection

Figure 1: Different analyses of the Jubileumsarkivet collection in Strix

2. Sparv

Språkbanken Text's analysis platform, Sparv, is a command-line tool for annotating text corpora [6]. The modular architecture of Sparv enables the integration of different corpus analyses on the token and sentence levels. Current analyses include tokanization, morphological, part-of-speech, compound, name entity recognition, sentiment, sentence segmentation, dependency parsing, and word sense disambiguation (WSD) [7], just to name a few.

Sparv is highly customizable due to its flexible plugin system. All modules (importers, annotators, exporters, model builders) are replaceable, making it easy to extend the system with custom plugins.[1]

[1] https://github.com/spraakbanken/sparv

3. Swedish frame semantic parser

Frame semantic parsing, the automatic processing of enhancing written text with semantic information, aims at identifying and classifying words and phrases according to their semantics following FrameNet principles [8] that are based on frame semantics theory [9, 10]. Training a frame semantic model requires access to large amounts of semantically annotated sentences, but with the advantages of recent transformer techniques, it was possible to train a language-specific parser on Swedish FrameNet [4], a resource with a relatively small amount of semantically annotated training data compared to English [5].

The parser is currently available in two model sizes—small and base [11] — both of which we retrained. Their evaluation results on the Swedish dataset are relatively low (see Table 1). Our goal is to train the large model, which is expected to yield significantly better performance.

Model	Task	Dataset	Precision	Recall	F1
Base	Args Extraction	Validation	0.501	0.450	0.474
	Frame Classification	Validation	0.468	0.468	0.468
	Trigger Identification	Validation	0.513	0.510	0.511
	Args Extraction	Test	0.521	0.456	0.487
	Frame Classification	Test	0.452	0.452	0.452
	Trigger Identification	Test	0.515	0.515	0.515
Small	Args Extraction	Validation	0.374	0.339	0.356
	Frame Classification	Validation	0.437	0.437	0.437
	Trigger Identification	Validation	0.454	0.451	0.452
	Args Extraction	Test	0.393	0.350	0.371
	Frame Classification	Test	0.434	0.434	0.434
	Trigger Identification	Test	0.419	0.408	0.413

Table 1
Comparison of Base and Small models on validation and test datasets for argument extraction, frame classification, and trigger identification.

Both models achieve moderate performance on the three evaluated tasks: argument extraction, frame classification, and trigger identification. In the validation set, the scores range from approximately 0.36 to 0.47 F1, with precision and recall showing balanced values. In the test set, the models achieve slightly higher performance in some cases, with F1 scores around 0.45–0.52. These results indicate that the models are able to capture relevant semantic and structural patterns in the data, though there is still room for improvement in terms of precision and recall, particularly for argument extraction. The larger model is expected to perform even better with competitive evaluation scores to English counterpart parsers.

4. Swedish Frame Semantic parser plugin

As part of the recent restructuring of Sparv, a plugin-based architecture was introduced that enables new annotation schemes and tools to be flexibly integrated into the processing pipeline. This modular design allows researchers to extend the system with minimal overhead, facilitating the addition of task-specific layers of linguistic analysis beyond the core annotations such as tokenization, POS tagging, and lemmatization.

Within this framework, we developed a plugin for frame-semantic parsing, aimed at providing semantic role information as an additional layer of text analysis. The plugin integrates the developed frame-semantic parser into the annotation pipeline, automatically identifying frames evoked by lexical units, and assigning frame elements to their corresponding arguments in the text. This enriches the corpus with structured semantic representations that go beyond surface-level syntactic analysis, enabling a more nuanced exploration of meaning in context. An example of how the plug-in operates is illustrated in the following.

Input
Ordförande för förhandlingarna har varit fröken M. Nordénfeldt, Göteborg.
Output
Activity: förhandlingarna
Leader: fröken M. Nordénfeldt
Place: Göteborg

In this example, the input sentence *Ordförande för förhandlingarna har varit fröken M. Nordénfeldt, Göteborg.* 'The chairman of the negotiations has been Miss M. Nordénfeldt, Gothenburg.' was parsed according to the semantic frame LEADERSHIP, trigged by the lexical unit *Ordförande* 'chairman'.

From a technical perspective, the plugin is designed to operate seamlessly within the standardized data exchange format of the infrastructure. It consumes pre-processed linguistic information (e.g., lemmas, POS tags, dependency relations) already available in the pipeline and outputs semantic annotations in a format consistent with other modules. This ensures interoperability with downstream components and facilitates combined analyses. The modular design further allows the plugin to be independently maintained or upgraded, ensuring its adaptability to future parser improvements or alternative semantic frameworks. The plugin is available as a github repository at https://github.com/shafqatvirk/sparv-plugin-semparse-swe.git, and the instructions to install and use a plugin can be found at https://spraakbanken.gu.se/sparv/user-manual/intro/.

5. Use case

The following use case illustrates a hypothetical example of how the frame-semantic parsing plugin could be applied to analyze diachronic Swedish texts and trace how the concept of *marknad* ('market') has shifted over time. With appropriate training or fine-tuning, the parser could distinguish between a Physical Market Frame, where *marknad* denotes a concrete site of exchange (involving elements such as Location, Seller, Buyer, Goods, Transaction, Time, and Atmosphere), and an Abstract Market Frame, where it refers to an economic or systemic domain (with elements such as Domain, Participants, Goods/Services, Mechanism, Condition, Competition, Outcome, and Influence). By applying these frame distinctions across historical corpora, we can quantify when and how references to *marknad* shift from physical to abstract contexts, thereby providing empirical evidence of semantic change in Swedish economic discourse.

6. Conclusion and future work

Sparv-SRL is a key plugin enabling advanced natural language understanding (NLU) for Swedish. Integrated into Språkbanken Text's annotation pipeline, it enriches sentences with semantic role information. The integration of the frame-semantic plugin demonstrates both the extensibility of the new infrastructure and the benefits of incorporating deeper linguistic analysis. By augmenting annotated corpora with semantic roles, the plugin opens new possibilities for linguistic research and downstream applications, including semantic search, discourse analysis, and studies of meaning variation over time.

In future work, we plan to enhance the search functionality in Strix, to enable researchers not only to analyze but also to visualize search results in ways that align with their specific research questions.

Acknowledgments

This work has been funded and supported by Språkbanken – jointly funded by its 10 partner institutions and the Swedish Research Council (2025–2028; project id 2023-00161) as well as (2018–2024; dnr 2017-00626). We would like to acknowledge the Swedish national research infrastructure Huminfra,

funded for the years 2022-2024 and 2005-2028, contracts 2021-00176 and 2023-00171 respectively, and the participating partner institutions.

References

[1] M. Forsberg, Y. Ali Mohammed, E. Sköldberg, M. Öhrman, SO in Strix: a lexicographic case study of entry vectors, in: Sixty years of Swedish computational lexicography / Dana Dannélls, Kristian Blensenius and Lars Borin (eds.), De Gruyter Brill, Berlin, 2025, pp. 289–303.

[2] M. Hammarstedt, A. Schumacher Olsson, L. Borin, M. Forsberg, Sparv 5.3.0: Språkbanken's Analysis Platform – Technical Report, Technical Report, University of Gothenburg, Göteborg, 2025.

[3] L. Borin, M. Forsberg, Saldo: the hub of Språkbanken's lexical research infrastructure, in: Sixty years of Swedish computational lexicography / Dana Dannélls, Kristian Blensenius and Lars Borin (eds.), De Gruyter, Berlin, 2025, p. 97–111.

[4] D. Dannélls, L. Borin, M. Forsberg, K. F. Heppin, M. T. Gronostaj, Swedish FrameNet, in: D. Dannélls, L. Borin, K. F. Heppin (Eds.), The Swedish FrameNet++: Harmonization, integration, method development and practical language technology applications, volume 14 of *Natural Language Processing*, John Benjamins, Amsterdam, 2021, pp. 37–66.

[5] D. Dannélls, R. Johansson, L. Y. Buhr, Transformer-based Swedish semantic role labeling through transfer learning, in: Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), ELRA and ICCL, Turin, Italy, 2024.

[6] M. Hammarstedt, A. Schumacher, L. Borin, M. Forsberg, Sparv 5 Developer's Guide, Technical Report, University of Gothenburg, 2022.

[7] R. Johansson, L. Nieto Piña, Combining relational and distributional knowledge for word sense disambiguation, in: Proceedings of the 20th Nordic Conference of Computational Linguistics (NODALIDA), volume 109, Linköping Electronic Conference Proceedings, Linköping University Electronic Press, Vilnius, Lithuania, 2015, pp. 69–78.

[8] J. Ruppenhofer, M. Ellsworth, M. R. Petruck, C. R. Johnson, C. F. Baker, J. Scheffczyk, FrameNet II: Extended Theory and Practice, Technical Report, 2016. URL: https://framenet2.icsi.berkeley.edu/docs/r1.7/book.pdf, berkeley: ICSI.

[9] C. J. Fillmore, Frame semantics and the nature of language, Annals of the New York Academy of Sciences 280 (1976) 20–32. URL: https://nyaspubs.onlinelibrary.wiley.com/doi/abs/10.1111/j.1749-6632.1976.tb25467.x. doi:https://doi.org/10.1111/j.1749-6632.1976.tb25467.x.

[10] C. J. Fillmore, Frame semantics, in: Linguistic Society of Korea (Ed.), Linguistics in the Morning Calm, Hanshin Publishing Co., Seoul, 1982, pp. 111–137.

[11] D. Chanin, Open-source frame semantic parsing, arXiv preprint arXiv:2303.12788 (2023).

Building a Vowel: a Bottom-up Guide for Phonetic and Language Learning Sciences

Axel Ekström[1,2,*], Runhui Song[2,3] and Jens Edlund[2]

[1]Centre for Cultural Evolution, Department of Psychology, Stockholm University, Stockholm, Sweden
[2]Speech, Music & Hearing, KTH Royal Institute of Technology, Stockholm, Sweden
[3]Department of Linguistics and Philology, Uppsala University, Uppsala, Sweden

Abstract

We present a guide for building a vowel "from the ground up" using software developed within the HumInfra infrastructure. Using this guide, a user may recreate in a simulation conditions for the production of a given vowel quality. It is summarized step-wise how a midsagittal view of a speaker's vocal tract may be reduced to a simplistic sequence of two-dimensional segments, which – input into the simulation software TubeN – predicts and recreates the given vowels quality. Our hands-on guide is of interest to students in the linguistic and teaching sciences, and to those teaching speech science in a broader sense.

Keywords

educational software, language learning, acoustics, phonetics

1. Introduction

Vowels arise from movements of articulators – including the jaw, lips, tongue, and velum. Such movements shape the acoustic signal in predictable ways by alternating the resultant spectral peaks, typically termed formants, generally held as the basis for vowel synthesis and perception [1]. The modern interpretation of this relationship stems from "source–filter theory" [1], which explains speech production as the interaction between a voice *source* (supplied through vocal fold oscillation) and a *filter* (the supralaryngeal vocal tract, beginning at the larynx and terminating and the mouth opening). While foundational in phonetic science and early speech synthesis, the framework has historically been studied within engineering disciplines [2, 3, 4, 5]. However, knowledge of speech production is not merely of use to those in speech sciences – but also to, for example, students of linguistics and language learning. Here, we present a step–by–step "ground-up" approach to constructing a simple, accessible model of vowel production, using attainable data and freely available software [6] developed within the *HumInfra* national infrastructure. This guide is intended for students in the broader humanities and educational sciences. While simplistic, the model yields predictable and replicable results, and it has already proven effective in workshops with phonetics students [7].

2. Building a vowel, step by step

2.1. Obtain vocal tract data

The goal of the present paper is a step-wise guide by which formant patterns observed in natural speech can be "reverse engineered" computationally, without an engineering background.[1] The first step toward such a model is data derived from articulation of the desired vowel – typically captured using magnetic resonance imaging (MRI) techniques. In Sweden, several universities now possess imaging equipment

HumInfra Conference 2025 (HiC 2025), Stockholm, 12–13 November 2025.
*Corresponding author.
axeleks@kth.se (A. Ekström)
0000-0002-6739-0838 (A. Ekström); 0000-0001-9327-9482 (J. Edlund)

[1]We do not describe nuances of formant estimation here, as sources toward this end are widely available elsewhere.

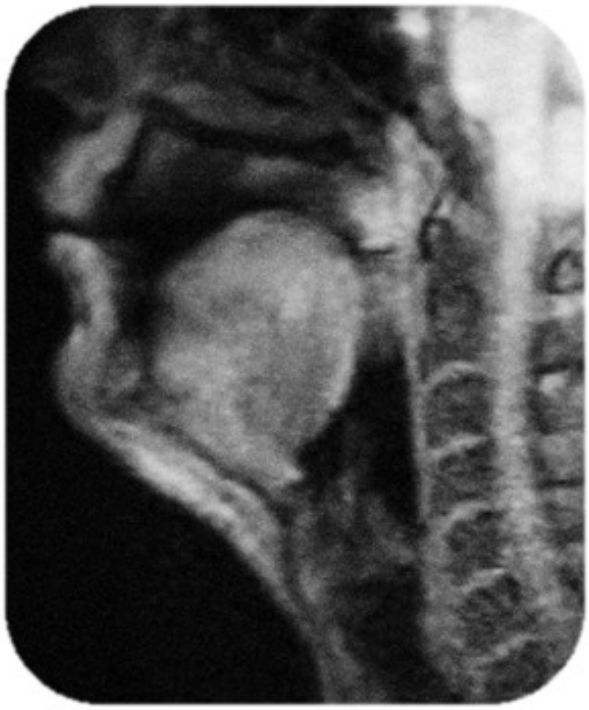


Figure 1: Adult male native Swedish speaker, aged 28, speaking [uː]. Data was collected during a magnetic resonance imaging (MRI) single-subject case study at the Stockholm University Brain Imaging Center. The scanner was a Siemens Prisma 3 Tesla whole-body MRI. Data collection used the "FLASH" sequence- a spoiled gradient echo with a flip angle of 5 degrees, an echo time of 1.7 ms, and a repetition time of 4 ms. The field of view was 280(freq) x 278(phase) mm, with a resolution of 224(freq) x 156(phase) and a bandwidth = 500 Hz / px.

necessary to record such data. These facilities include the Stockholm University Brain Imaging Center (SUBIC) and the Lund University Bioimaging Center. However, note that databases of such data are also available [8, 9]; as such, it is not necessary that new data be acquired. For the sake of illustration, we selected from data previously recorded for other research purposes, an occurrence at which the speaker (an adult male) produced long close back rounded vowel [uː] (Figure 1).

2.2. Segmentation

A midsagittal view of a speaker's vocal tract at the moment of vowel realization can be simplistically reduced to a sequence of smaller segments, defined by their length and area. For [uː] – the above stated test case – an initial *narrow* section (the rounded and protruded lips) is followed by an expansive *open* section (the anterior oral cavity, where the tongue has been retracted, leaving more open space), etc. To achieve useable segment lengths and areas, the vocal tract shape may be traced and isolated from its surrounding anatomy. A line may then be traced at the midpoint of the distance from one wall to the other (up-to-down, or left-to-right, depending on the position in the tract). By then tracing equidistant sections in the structure, segments defined by their *distance in the sagittal plane* may be derived for computational implementation. In theory, the lengths of such segments may be varied. However, both 1cm and 0.5cm segment lengths are commonplace in the literature.

2.3. Converting distances to areas

Articulatory-acoustic dimension reduction techniques developed over several decades [3, 1, 10] have illustrated that the vocal tract – a three-dimensional complex of biological elements – can be simplistically modeled as a two-dimensional sequence of segments defined by their length and area. Without direct

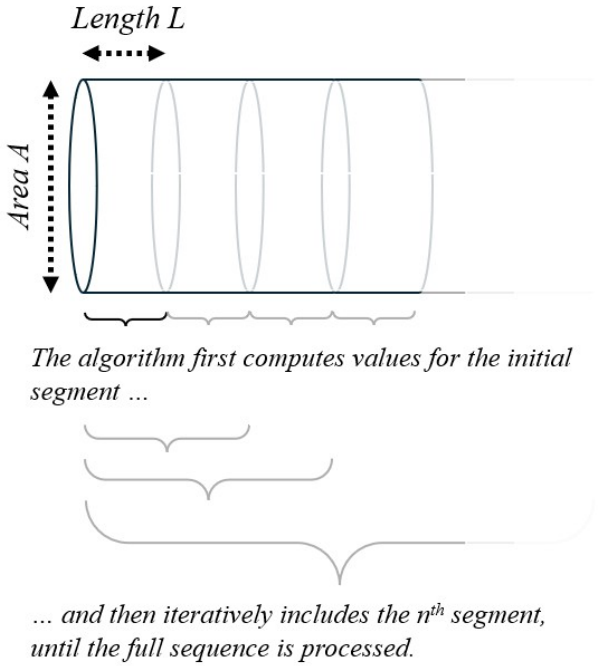


Figure 2: Workflow of the Liljencrants and Fant algorithm, as implemented in *TubeN*. The program makes no restrictions on the number or lengths of the total sequence.

access to three-dimensional vocal tract data – which requires more invasive, and more specialized methodology – it is necessary to convert cross-sectional relationships to area estimates for each segment. Fant [11] stipulates such an equation which is appropriate for the present purposes, according to which conversions from distance *d(x)* in the sagittal plane to cross-sectional area *A(x)* may be computed as

$$A(x) = a \cdot d(x)^b \tag{1}$$

Specifically, however, because this relationship is not constant throughout the vocal tract, it is necessary to adapt the power function based on what part of the image is currently being processed.

For the lip section, Fant [11] stipulates values $a = 1.8$ and $b = 2.5$ where $d > 1.7$ cm, and $a = 1.8$ and $b = 2.5$ where $d < 1.7$ cm.

For the volume of the oral cavity, Fant [11] suggests including corrections for air columns on both sides of the tongue, which is retracted during sustained production of [u:]. As such, the user may implement $a = 2.4$ and $b = 1.4$, before adding a correction of an additional 35% to the final volume [11].

Finally, Fant [11] notes that the power function should be subdivided for application to the pharynx, as its midsagittal distance-to-volume relationships varies significantly. He suggests that where $d < 1.75$, $a = 2$ and $b = 1.6$; and where $1.75 < d < 2.5$, $a = 2.8$ and $b = 1$.

2.4. Computing the properties of an acoustic tube

The TubeN software [6] implements an algorithm developed by Liljencrants and Fant [12], which predicts, for any sequence of tube segments, the resultant spectral peaks (functionally equivalent to formants).[2] That is, if completed appropriately, formants predicted by TubeN for a shape corresponding to the speaker's vowel production, should closely match the vowel quality produced by that speaker at that moment. The TubeN software is publicly available and can be accessed online through the relevant GitHub repository.[3]

[2]For the sake of brevity, the mathematical representation of the program is not included here. Readers are directed to the original publication [12, 6].
[3]https://github.com/jbeskow/tuben

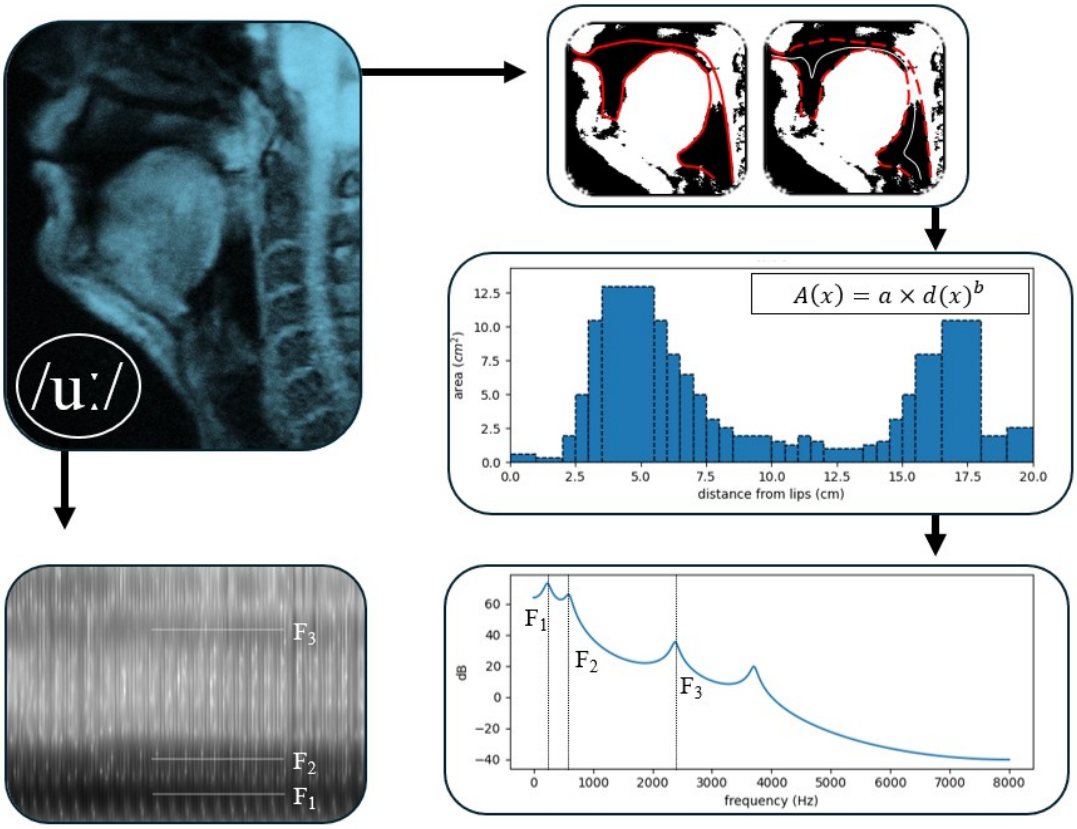


Figure 3: Note that segments are not annotated. In theory, segments can be of variable length and – provided relationships between them are mostly preserved – predicted vowel qualities will overlap approximately.

3. Concluding comments

In brief, the "vowel reverse engineering" exercise described above involves four steps. They are, in order, (1) obtain or collect midsagittal view of a speaker vocal tract mid vowel production; (2) trace the effective tract, from glottis to lips; (3) segment said outline into equidistant "slices" and (4) convert them into appropriate cross-sectional area estimates; (5) input lengths and areas of segments into the *TubeN* software, which automatically generates predicted resonant frequencies (or formants); and finally, (6) match with formants estimated during live production of the same vowel. We look forward to exploring the potential of this procedure as a teaching tool in broader phonetics and language pedagogical education, and are open to collaborations with any interested parties toward this end.

Acknowledgments

The results of this work and the tools used will be made more widely accessible through the national infrastructure Språkbanken Tal under funding from the Swedish Research Council (2017–00626). We thank Gláucia Laís Salomão and Jonathan Berrebi at the Stockholm University Brain Imaging Center for their assistance in recording our MRI data. We extend additional thanks to Jonas Beskow for discussion about the program, and to the attendants of the Swedish Linguistics (SLING) meeting of 2023, and the Fonetik 2024 Conference, where the software and its application was extensively discussed.

References

[1] G. Fant, Acoustic Theory of Speech Production: With Calculations Based on X-Ray Studies of Russian Articulations, Mouton, 1971.

[2] B. E. Lindblom, J. E. Sundberg, Acoustical consequences of lip, tongue, jaw, and larynx movement, The Journal of the Acoustical Society of America 50 (1971) 1166–1179.
[3] H. K. Dunn, The calculation of vowel resonances, and an electrical vocal tract, The Journal of the Acoustical Society of America 22 (1950) 740–753.
[4] P. Badin, G. Fant, Notes on vocal tract computation, STL QPSR 2 (1984) 53–108.
[5] K. N. Stevens, S. Kasowski, C. G. M. Fant, An electrical analog of the vocal tract, The Journal of the Acoustical Society of America 25 (1953) 734–742.
[6] R. Song, J. Beskow, J. Edlund, M. Tronnier, R. Tu, K. Zhang, A. Ekström, Open source software for tube vocal tract modeling, resonance prediction, illustration, and 3D printing, BioRXiv (2025). doi:https://doi.org/10.1101/2025.10.15.682256.
[7] M. Tronnier, A. G. Ekström, Teaching speech acoustics through vocal tract modeling, in: G. Ambrazaitis, Raschellà, N. J. Young (Eds.), Proceedings from FONETIK 2025, Linnaeus University, 2025, pp. 83–84.
[8] T. Sorensen, Z. I. Skordilis, A. Toutios, Y. C. Kim, Y. Zhu, J. Kim, S. S. Narayanan, Database of volumetric and real-time vocal tract mri for speech science, in: Proceedings of Interspeech, 2017, pp. 645–649.
[9] Y. Lim, A. Toutios, Y. Bliesener, et al., A multispeaker dataset of raw and reconstructed speech production real-time mri video and 3d volumetric images, Scientific Data 8 (2021) 187. doi:10.1038/s41597-021-00976-x.
[10] R. Carré, P. Divenyi, M. Mrayati, Speech: A dynamic process, Walter de Gruyter GmbH & Co KG, 2017.
[11] G. Fant, Vocal tract area functions of swedish vowels and a new three-parameter model, in: Proceedings of the 2nd International Conference on Spoken Language Processing (ICSLP 1992), ISCA, 1992, pp. 807–810. doi:10.21437/ICSLP.1992-262.
[12] J. Liljencrants, G. Fant, Computer program for vt-resonance frequency calculations, STL-QPSR 16 (1975) 15–21.

Beyond Big Tech: Alternative Digital Platforms for Collaborative and Participatory Art Historical Research

Elizabeth Ashley Fox-Jensen[1]

[1] *Malmö University, Faculty of Culture and Society, Sweden*

Abstract

The selection of digital collaboration platforms impacts research participation in international digital humanities projects. This study emerged from practical challenges during the "Ted Stamm: *Tags*" project, a multi-institutional art historical research initiative transcribing 63 sketchbooks (1973–81) with 675 documented participant contributions. Initial reliance on Google Sheets was discontinued due to ethical concerns regarding policy changes, while the subsequent transition to Microsoft Excel created barriers for external collaborators across different institutional frameworks.

This paper investigates alternative collaborative platforms that meet European standards for data sovereignty while supporting multi-institutional research collaboration. The research question asks: What European alternative platforms exist that provide institutional compatibility and GDPR compliance without sacrificing collaborative functionality? Through a case study methodology grounded in *Tags* transcription project, this paper proposes an evaluation structure and planned comparative assessment of three European platforms: kSuite, LibreOffice, and Proton Drive.

The evaluation framework assesses platforms across six criteria: (1) institutional compatibility with external collaborators, (2) real-time collaboration capabilities, (3) data sovereignty and GDPR compliance, (4) scalability for research projects, (5) integration with existing academic workflows, and (6) cost sustainability. This research responds to a need identified in Huminfra infrastructure guidance by systematically evaluating EU alternatives to US-based platforms.

The paper is structured as follows: Section 1 introduces the research context and platforms. Section 2 examines European digital sovereignty. Section 3 reviews Ted Stamm's artistic practice and the *Tags* case study. Section 4 develops the evaluation framework with testing protocol. Section 5 discusses preliminary platform assessment findings. Section 6 concludes recommendations for digital humanities and institutional research.

Keywords

Digital sovereignty, collaborative platforms, participatory design, participatory art, art history, conceptual, European research infrastructure, GDPR, digital humanities, Ted Stamm

1. Introduction

Digital tool selection determines the participants in international research collaborations. The 2025 transition within our *Tags* research team from Google Sheets to Microsoft Excel, initially an ethical response to policy changes, revealed systematic barriers that excluded qualified researchers based on institutional IT configurations rather than scholarly merit [1].

These challenges reflect broader European movements toward data sovereignty, as exemplified by policy shifts across Denmark, Germany, and Sweden [2] [3] [4]. When institutional IT policies create access delays for external collaborators, platform selection becomes a methodological and ethical imperative. The current situation is complicated by heightened dependency on US-owned platforms: specifically, Microsoft and Google, whose recent policy shifts have minimized or ended DEI initiatives and GDPR protections.

This study focuses specifically on institutional compatibility challenges, authentication barriers, and collaborative cross-border workflows. While we acknowledge broader accessibility considerations including assistive technology compatibility and universal design principles, this research prioritizes platform evaluation for multi-institutional research coordination rather than focusing on comprehensive accessibility auditing. Future research should systematically examine the compliance of these platforms

Huminfra Conference 2025, Stockholm, 12-13 November 2025.

elizabeth.jensen@mau.se (Elizabeth Ashley Fox-Jensen) https://orcid.org/0000-0003-2004-8039

with WCAG 2.1 standards and compatibility with screen readers, cognitive support tools, and mobile accessibility features.

This research evaluates readymade commercial services rather than self-hosted solutions such as Nextcloud [5]. Offering maximum institutional control, self-hosting nonetheless requires substantial IT infrastructure, specialized technical expertise, and resources for ongoing maintenance that exceed the capacity of most humanities research projects. Commercial EU alternatives provide GDPR compliance and data sovereignty without requiring in-house server administration, making them more feasible for typical digital humanities (DH) collaborations.

2. European Digital Sovereignty and Platform Independence

European governmental and academic stakeholders increasingly recognize that dependence on US-owned proprietary platforms undermines institutional independence and legal compliance under GDPR and the Digital Markets Act [6]. Denmark's Digital Minister Caroline Stage (2025) [7] articulated this concern: "Too much public digital infrastructure is tied up with very few foreign suppliers," creating vulnerabilities extending beyond economics to encompass data security, research autonomy, and democratic governance of academic infrastructure.

Institutional transitions show feasibility at scale. The German state of Schleswig-Holstein's migrated sixty thousand government workers to LibreOffice and Linux-based systems, providing a template with documented benefits including reduced costs, improved security, and GDPR compliance [8]. These developments, however, do not represent coordinated national strategies. While focusing on humanities research infrastructure coordination, Huminfra does not propose specific alternatives to US-based platforms [9], bringing awareness to the need for systematic assessment frameworks.

The EU's Digital Decade strategy emphasizes sovereign digital capabilities that compete with global technology monopolies and maintain European values of privacy, transparency, and social responsibility [10]. For humanities research, this translates to infrastructure supporting diverse methodological approaches, multilingual collaboration, and cross-border participation within regulatory compliance frameworks.

3. Ted Stamm and the Participatory Tags Project: Case Study

3.1 Ted Stamm and the *Tag* Collaborative Sketchbooks

Ted Stamm (1944–1984) was a conceptual artist whose practice emerged from the artistic context of the 1970s SoHo neighborhood in New York. A graduate of Hofstra University (1967), where he studied under the painter Perle Fine, Stamm developed a distinctive artistic methodology centered on relinquishing personal control through strategic use of chance, found materials, and collaborative participatory based processes. His artistic evolution reflects sustained engagement with questions of artistic authorship, democratic participation, and the artist's role in the creative process.

Beginning in 1972, Stamm moved away from lyrical Abstract Expressionism toward a more conceptually rigorous practice. He developed his *Cancel* series by systematically covering existing paintings with dense arrays of marks in varying shades of black, embracing destruction as a form for artistic renewal. He concurrently began to collect tags with paper strings that had been discarded on the streets of SoHo: industrial remnants of the neighborhood's warehouses and manufacturing centers. These found tags became the foundation for a decade-long experimental practice using chance, mark making, and participatory artistic dialogue.

In the *Tag* Collaborative Sketchbooks (1974–81) Stamm explored participatory artistic practice. Rather than presenting himself as the sole artistic authority, he invited studio visitors and individuals he randomly encountered outside his studio to participate in collaborative dialogues of mark making. The format was deceptively simple yet conceptually profound: Stamm created matching pairs of Sennelier French spiral notepads, affixing one found garment tag to each page. In the first notebook, labeled "Executed by Individuals" (Figure 1), participants responded to prompts such as "Price a tag," "Blacken a tag," or "Draw a horizontal line through a tag." In the second notebook, labeled "Artist

Book," Stamm responded to these marks with his own interventions. Both pages were stamped with dates and other documentation to create permanent records of these participatory based works.

Stamm's participatory practice shows deliberate commitment to inclusive practices and democratic artistic engagement. His documented collaborators span diverse constituencies, ranging from prominent art world figures such as William Zimmer (art critic at the *New York Times*) and Heidi Colsman-Freyberger (executive director of Barnett Newman Foundation) to studio visitors and people encountered purely by chance. This inclusive approach questioned traditional hierarchies of artistic authority and authorship, positioning Stamm simultaneously as an artist, a facilitator, and a responsive participant in collective creative processes.

The materiality of Stamm's participatory work merits particular attention. Participants employed graphite pencils, paint, spray paint, markers, rubber stamps, foam brushes, staples, sandpaper, and even their own thumbprints (Figure 2), materials that reflected diverse gestures and mark-making vocabularies. The garment tags themselves became conceptually charged objects, transforming industrial detritus into a site of artistic dialogue. Some collaborators embraced delicate gestures; others made bold, messy or decisive marks. Stamm's responses acknowledged and amplified these varied interventions, creating dialogues in which his artistic judgment served as collaborative exploration rather than hierarchical control.

The *Tag* series achieved international recognition in Documenta 6 (Kassel, Germany, June 24–October 2, 1977) [11], where curator Manfred Schneckenburger featured Stamm's participatory *Tag* notebooks in the exhibition's "Art Books" section. This recognition situated Stamm's participatory methodology within broader international discourse on conceptual and experimental artistic practices, particularly in relation to questions of authorship, process-based art, and audience engagement that dominated the artistic discourse of the 1970s.

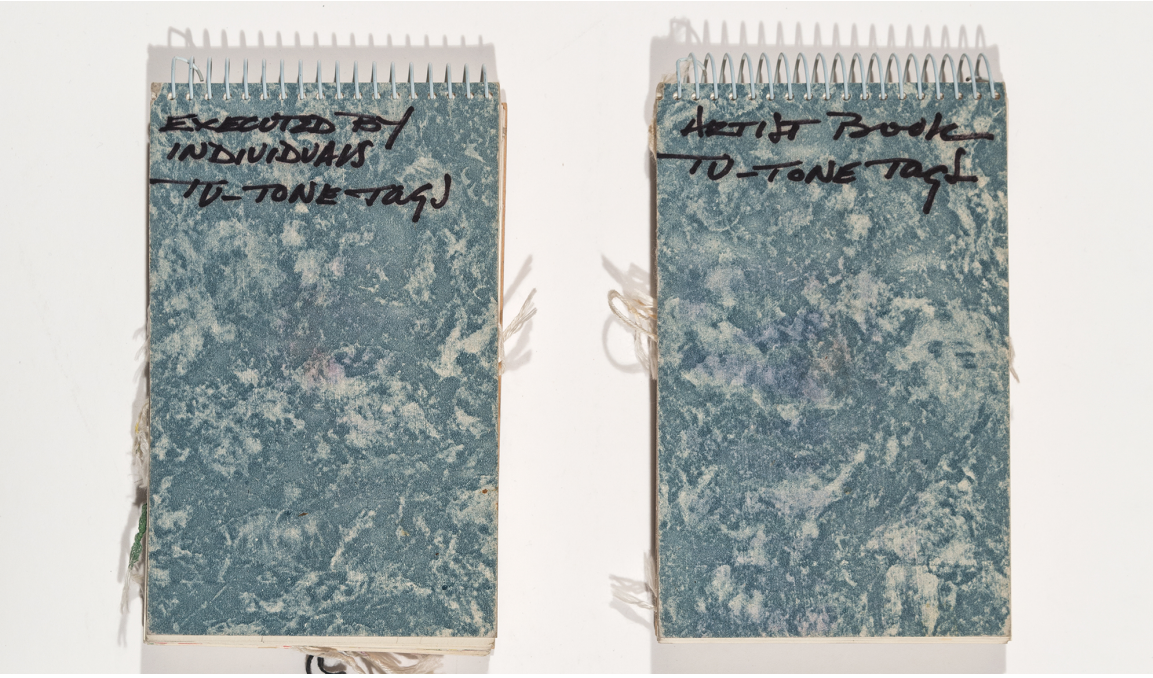


Figure 1. Ted Stamm, *Tags Sketchbook* 61 (1977). Left: "Executed by Individuals Tu-Tone Tags" collaborative works by twenty-five participants, illustrating the project's participatory method. Right: Ted Stamm's original composition, "Artist Book Tu-Tone Tags."

Figure 2. Ink Thumbprint Tags from Ted Stamm, *Tags Sketchbook* 80 (1980). Left: Original page detail. Right: Contribution by Per Haubro Jensen (May 1980), showing individual intervention in the collective process.

3.2 The Tags Transcription Project: Research Context

The Ted Stamm *Tags* transcription project coordinated collaborative research across 63 sketchbooks containing approximately 1,575 pages of artistic documentation that span the eight-year collaborative period (1973–81). This multi-institutional initiative assembled research fellows from Malmö University, the University of Chicago, Yale University, School of Advanced Study (SAS) at the University of London, and Rice University, supported by the Jeff Metcalf Internship Program and Arts and Humanities Bridge initiative at the University of Chicago.

The project's scale and scope, documenting approximately 675 identified contributors to Stamm's participatory practice, required systematic transcription cataloguing of detailed metadata including contributor names, collaboration dates, materials employed (paint, graphite, spray paint, markers, staples, hole punches, stamps), and analytical descriptions of individual contributions within collective processes. This comprehensive documentation created a dataset suitable for multiple analytical approaches including art historical contextualization within 1970s conceptual art practices, urban history analysis examining the SoHo neighborhood as context and medium, and mapping of participatory chronology examining participant backgrounds throughout the collaborative period.

Research fellows conducted intensive transcription work in July–August 2025, developing standardized metadata templates and quality control protocols. The systematic approach ensured scholarly rigor that supported multiple potential research outputs: peer-reviewed publications examining participatory artistic methodologies, art historical essays contextualizing the *Tags* series within broader artistic movements, and digital humanities (DH) projects exploring collaborative practices through archival data visualization.

3.3 Institutional Compatibility Barriers

Challenges emerged at the intersection of technical limitations and institutional security protocols. GDPR-compliant data handling at Malmö University, while legally essential, created unexpected barriers when external collaborators could not access shared Microsoft Excel spreadsheets (Figure 3) through institutional SharePoint, effectively excluding researchers based on institutional affiliation rather than scholarly qualifications [12]. These delays particularly affected international collaborators from institutions with differing IT security protocols.

The solution involved transitioning to School of Advanced Study (SAS) at the University of London's Microsoft SharePoint, configured to facilitate external institutional participation. This work-around succeeded and highlighted systemic dependence on proprietary Microsoft infrastructure with varying institutional configurations.

Beyond policy and access barriers, the research team encountered specific technical limitations. Microsoft Excel's absence of native multiuser real-time editing (without SharePoint integration) required sequential workflow modifications that slowed transcription progress. Google Sheets' formula compatibility issues when exporting to Excel formats required manual data verification that created additional quality control burdens. SharePoint's complex permission hierarchies required IT administrator intervention for each new external collaborator, creating one-to-two-week delays for access provisioning. These technical friction points, combined with ethical and legal concerns, strengthened the imperative for evaluating European alternatives.

	A	B	C	D	E	F	G	H	I
1	**SK**	**Page**	**A (Stamm)/B (others)**	**Date**	**Image file name**	**Size (in.)**	**Participant**	**Description**	**Check Participant**
1196	80	8	A	5/11/1980	TS_Tag-Colab_SK_080_Thumbprint_1980_DSC9677	6 x 3.75	Per Haubro Jensen	Tag adhered upside down, date in righthand bottom corner, dark thumbprint placed slightly at a left angle	
1197	80	8	B	5/11/1980	TS_Tag-Colab_SK_080_Thumbprint_1980_DSC9677	6 x 3.75	Ted Stamm	Tag adhered upside down, narrow/ partial thumbprint in center of tag, signature large in bottom right	
1198	80	9	A	5/11/1980	TS_Tag-Colab_SK_080_Thumbprint_1980_DSC9678	6 x 3.75	Peter Fend	Tag adhered upside down, date in righthand bottom corner, dark thumbprint over tag	
1199	80	9	B	5/11/1980	TS_Tag-Colab_SK_080_Thumbprint_1980_DSC9678	6 x 3.75	Ted Stamm	Tag adhered upside down, date in righthand bottom corner, thumbprint applied horizontally over tag and part of string in lower 3/4 of tag	
1200	80	10	A	5/11/1980	TS_Tag-Colab_SK_080_Thumbprint_1980_DSC9679	6 x 3.75	Geneene Estrada	Tag adhered upside down, date in righthand bottom corner, thumbprint applied horizontally over tag and part of string in lower 3/4 of tag	
1201	80	10	B	5/11/1980	TS_Tag-Colab_SK_080_Thumbprint_1980_DSC9679	6 x 3.75	Ted Stamm	Tag adhered upside down, two partial separated thumbprints, one darker applied over tag and string arranged over tag. "Maybe May"	
1202	80	11	A	5/11/1980	TS_Tag-Colab_SK_080_Thumbprint_1980_DSC9680	6 x 3.75	John Ford	Tag adhered upside down, date in righthand bottom corner, this time along vertical edge, dark thumprint applied over string and tag, split into two	
1203	80	11	B	5/11/1980	TS_Tag-Colab_SK_080_Thumbprint_1980_DSC9680	6 x 3.75	Ted Stamm	Tag adhered upside down, dakr thumbprint applied over tag and part of string and swept upward to the right	
1204	80	12	A	5/11/1980	TS_Tag-Colab_SK_080_Thumbprint_1980_DSC9681	6 x 3.75	Don Hazlitt	Tag adhered upside down, date in righthand bottom corner along vertical edge, thumbprint applied over top half of tag and spills onto sketchbook	
1205	80	12	B	5/11/1980	TS_Tag-Colab_SK_080_Thumbprint_1980_DSC9681	6 x 3.75	Ted Stamm	Tag adhered upside down, lightweight thumbprint over tag	
1206	80	13	A	5/12/1980	TS_Tag-Colab_SK_080_Thumbprint_1980_DSC9682	6 x 3.75	J. Handzel	Tag adhered upside down, date in righthand bottom corner along vertical edge, thumbprting applied over tag in lighter ink application	
1207	80	13	B	5/12/1980	TS_Tag-Colab_SK_080_Thumbprint_1980_DSC9682	6 x 3.75	Ted Stamm	Tag adhered upside down, dark thumbprint applied to the left of tag over strings	

Figure 3. Transcription spreadsheet for Ink Thumbprint Tags from *Tags Sketchbook* (80), capturing contributor metadata (e.g., Peter Fend, Geneene Estrade, John Ford, Per Haubro Jensen, Don Hazlitt, and J. Hendzel).

4. Evaluation Framework and Testing Protocol

To address systematic challenges, this research proposes a framework centered on six criteria for collaborative humanities research [13] [14]: (1) Institutional Compatibility, (2) Real-Time Collaboration, (3) Data Sovereignty, (4) Scalability, (5) Integration, and (6) Cost Sustainability.

4.1 Assessment Criteria (1–5 Scale)

Each platform receives a score from 1 (Poor) to 5 (Excellent) based on practical testing:

1. Institutional Compatibility
    - Testing: External collaborator access without institutional accounts; guest protocols
    - Target: Seamless access for researchers from multiple universities
2. Real-Time Collaboration
    - Testing: Multiple simultaneous editors; tracking and version history; synch reliability
    - Target: Google Sheets equivalent functionality without proprietary lock-in
3. Data Sovereignty
    - Testing: Server locations (European-based); GDPR compliance documentation; data export/deletion capabilities; and terms of service review
    - Target: Full compliance with European data protection standards
4. Scalability

- Testing: Performance with increasing users; storage limits; large file handling; and project growth accommodation
- Target: Support for research projects ranging from small teams to large collaborations

5. Integration
   - Testing: Excel/CSV import-export; compatibility with citation managers; API availability
   - Target: Minimal workflow disruption when transitioning from existing platforms
6. Cost Sustainability
   - Testing: Pricing structures for academic projects; long-term affordability; free tier limitations; and institutional licensing options
   - Target: Budget-feasible alternatives to Microsoft 365/Google Workspace

4.2 Platform Selection Rationale

kSuite by Infomaniak [15] (Switzerland): Comprehensive collaboration suite including kDrive (cloud storage), online office tools, and communication features. Swiss GDPR compliance and data sovereignty guarantees. Pricing: €1.76–€6.92 user, month, with free tier available.

LibreOffice [16] (Germany): Open-source office suite offering full Microsoft Office compatibility without vendor lock-in. Cross-platform with strong international community support. Cost: Free. Note: Requires additional cloud storage solution for collaboration (can integrate with Nextcloud, ownCloud, or other platforms).

Proton Drive [17] (Switzerland): End-to-end encrypted cloud storage with emerging collaboration features. Swiss GDPR compliance. Pricing: Free (5GB); €4.99 per month (200GB). Note: Collaboration features are currently limited compared to Google Drive / Microsoft SharePoint.

4.3 Two-Week Testing Protocol

This streamlined approach prioritizes real-world usability testing with researchers facing barriers.

Target Test Groups:

Group 1: External institutional researcher with different IT systems (1 participant)

Group 2: International researcher under different GDPR frameworks (1 participant)

Weeks 1–2: User Testing

Real collaborative tasks: Shared spreadsheet editing, document coauthoring, data entry, and file version management. Feedback collection via surveys and semistructured interviews.

Week 3: Assessment

Score each platform (1–5) across six criteria based on user feedback and technical testing. Cost analysis and institutional IT requirement verification.

Week 4: Analysis and Recommendations

Compile results and develop practical recommendations for conference presentation, focusing on which platforms suit different research collaboration scenarios.

Data Collection: Postsession feedback and interviews with participants

5. Preliminary Platform Assessment

Based on available documentation and initial pilot testing, preliminary findings suggest:

Platform Strengths:

kSuite by Infomaniak:

- Strong institutional compatibility (4/5), familiar interface reduces training, multilingual support, comprehensive collab tools, Swiss data sovereignty (5/5), and competitive pricing

LibreOffice:

- Maximum cost sustainability (free/5), strong offline functionality (5/5), excellent Microsoft compatibility (4/5), no vendor lock-in, and large open-source community support

Proton Drive:

- Maximum data sovereignty and encryption (5/5), Swiss privacy protection, strong security model, emerging collaboration features, and affordability

Identified Challenges:

- Institutional Integration: European platforms currently score 2–4/5 versus US platforms' 3–5/5 for seamless institutional SSO (single sign-on) integration
- Real-Time Collaboration: LibreOffice requires additional cloud infrastructure; Proton Drive's collaboration features are still developing as compared to Google Drive maturity
- Learning Curve: Higher initial training investment required compared to familiar Google / Microsoft interfaces
- Feature Parity: Some European alternatives still developing feature completeness for advanced collaboration workflows

Cost Analysis:

LibreOffice offers unmatched cost sustainability as free open-source software. kSuite by Infomaniak and Proton Drive provide competitive pricing (€1.76–€6.92 per month and €4.99 per month, respectively) compared to Microsoft 365 (€10.50 per month) and Google Workspace (€5.75–€15.60 per month). Analysis of the total cost of ownership should account for institutional IT support requirements, staff training time (estimated 8–16 hours per researcher), and workflow modification investments beyond license costs. Organizations transitioning from established platforms should budget 15–20% additional overhead during migration phases.

6. Conclusions and Recommendations

6.1 Key Findings

European alternative platforms offer viable options for multi-institutional DH collaboration while meeting data sovereignty requirements. Success requires balancing technical functionality, institutional compatibility, and regulatory compliance. No single platform provides perfect feature parity with Google / Microsoft ecosystems, but combinations of these platforms can address specific research needs. The experience with the *Tags* project shows that platform selection impacts research participation equity, with institutional access barriers creating systematic exclusion unrelated to scholarly qualifications.

6.2 Institutional Recommendations

Research Projects

Research teams should implement processes of participatory platform selection that involve all collaborators from project inception. This participatory approach ensures that platform choices such as kSuite by Infomaniak, LibreOffice, or Proton Drive reflect the needs and constraints of all team members rather than institutional convenience or technical preferences. Institutions should allocate resources for training and transition support when moving to new platforms, recognizing that the learning curve for unfamiliar interfaces represents an implementation cost.

Beyond initial transition planning, research projects require clearly established contingency protocols for addressing access barriers as they emerge, ensuring that systematic inequities do not exclude qualified collaborators mid-project. Throughout the implementation process, research teams should systematically document platform limitations and compatibility challenges they encounter, creating documentation that can inform institutional infrastructure advocacy and support broader efforts to develop European digital alternatives.

IT Departments

Institutional IT departments must develop guest access protocols explicitly designed to accommodate European platforms alongside existing proprietary systems. This requires moving beyond security policies designed exclusively for Microsoft and Google ecosystems to create authentication procedures compatible with kSuite by Infomaniak, LibreOffice, and Proton Drive. IT departments should designate technical support personnel with expertise in authentication services and cross-institutional log-in

protocols, recognizing that their support needs differ substantially from conventional enterprise platforms.

Additionally, IT departments should undertake comprehensive security policy reviews to identify unnecessary barriers that prevent the adoption of GDPR-compliant alternatives, distinguishing between security requirements and legacy policies designed around proprietary platform assumptions. Finally, institutions should investigate institutional site licenses and volume pricing arrangements with European collaboration tool providers such as kSuite by Infomaniak, LibreOffice, and Proton Drive, potentially reducing per-user costs and facilitating broader adoption across multiple departments and research initiatives.

Policy Development

Institutional and governmental policy frameworks must establish clear guidelines that balance legitimate security requirements with the collaborative accessibility necessary for inclusive research environments. These guidelines should explicitly recognize that excessive security restrictions can constitute barriers to participation, requiring careful calibration between protection and accessibility. Policy development requires sustained coordination among legal compliance, IT security, and research leadership to guarantee that data protection regulations do not inadvertently exclude collaborators or enable the reproduction of institutionalized big tech inequalities.

National and institutional policies should actively support European digital sovereignty through dedicated infrastructure investment and targeted funding for platforms meeting GDPR and European privacy standards. Finally, policy development should prioritize establishing best practices for multi-institutional cross-border collaboration, documenting effective protocols for external guest access, international researcher participation, and navigation of diverse regulatory frameworks within single research initiatives.

6.3 Future Research

Longitudinal studies tracking platform adoption outcomes, including participation patterns, research productivity, and total cost of ownership across different project scales must be undertaken. Systematic evaluation of emerging European platforms as collaboration features mature. Furthermore, comparative analysis of self-hosted versus commercial European solutions should be framed by institutions with IT infrastructure capacity.

Also, investigation of innovative cross-border collaboration models are needed to show how European platforms facilitate international partnerships under diverse regulatory frameworks. Analysis of legal standards, technical protocols, and cultural practices supporting or hindering inclusive collaboration might be examined.

Exploration of participatory design methodologies for platform selection would help examine the influence of researcher input on institutional technology decisions.

6.4 Broader Implications

Platform selection in humanities research constitutes ethical practice extending beyond technical functionality to questions of institutional autonomy, research independence, and democratic knowledge creation. European alternative platforms show that data sovereignty and collaborative effectiveness represent complementary rather than competing goals. Success requires sustained institutional commitment, meaningful community engagement, systematic resource allocation, and recognition that digital tools inform the knowledge created and communities engaged in its production. Forward-looking research infrastructure must directly embody scholarly values of openness and intellectual freedom within design and implementation processes.

The Ted Stamm *Tags* project exemplifies this principle: Stamm's deliberate choice to include diverse voices in collaborative artistic practice mirrors the needs of research communities for inclusive, equitable digital infrastructure supporting meaningful participation across institutions.

Acknowledgments

This research emerged from the "Ted Stamm: *Tags*" project conducted in collaboration with the Ted Stamm Estate and researchers from Malmö University, the University of Chicago, Yale University,

School of Advanced Study (SAS) at the University of London, and Rice University, supported by the Jeff Metcalf Internship Program and Arts and Humanities Bridge initiative at the University of Chicago. Special thanks to the institutional IT departments for platform testing support, the editor Amelia Kutschbach, and the Huminfra 2025 conference organizers.

References

[1] Fox-Jensen, E. A. (2025). In-progress, "Ted Stamm: *Tags*" participatory based works. Malmö University, Sweden.
[2] Stage, C. (2025). Denmark's Digital Policy Announcements. Ministry of Digital Affairs. https://www.english.digmin.dk/the-minister.
[3] Juul, J. (2025). *The Ludologist: European Alternatives to Google and Microsoft*. https://www.jesperjuul.net/ludologist.
[4] 2Data (2025). *A Search for Digital Sovereignty: EU Governments Shift from Microsoft to Linux and LibreOffice*. https://www.2-data.com/knowledge-hub/a-search-for-digital-sovereignty-eu-governments-shift-from-microsoft-to-linux-libreoffice.
[5] Nextcloud (2025). *Nextcloud—Open-Source Content Collaboration Platform*. https://nextcloud.com.
[6] BBC (2025). *Should Europe Wean Itself off US Tech?* https://www.bbc.com/news/articles/c3dpr2zkny0o.
[7] Stage, C. (2025). Denmark's Digital Policy Announcements.
[8] 2Data (2025). A Search for Digital Sovereignty: EU Governments Shift from Microsoft to Linux and LibreOffice.
[9] Huminfra/Lund University (2025). National Research Infrastructures. https://www.humlab.lu.se/research/national-research-infrastructures.
[10] ALLEA (2014). *Facing the Future: European Research Infrastructures for Research and Innovation.* All European Academies.
[11] Saletnik, J. (2023). Ted Stamm's 1972 Reset. In *Ted Stamm: Series* (pp. 82–89, 231–33). Hatje Cantz.
[12] Fox-Jensen, E. A. (2025). In-progress, "Ted Stamm: *Tags*."
[13] ALLEA (All European Academies) Working Group E-Humanities (2023). *Recognising Digital Scholarly Outputs in the Humanities*. ALLEA. https://allea.org/portfolio-item/recognising-digital-scholarly-outputs-in-the-humanities.
[14] American Historical Association (2015). *Guidelines for Professional Evaluation of Digital Scholarship by Historians*. https://www.historians.org.
[15] Infomaniak (2025). *kSuite*. https://www.infomaniak.com/en/ksuite.
[16] The Document Foundation (2025). *LibreOffice*. https://www.libreoffice.org.
[17] Proton AG (2025). *Proton Drive*. https://proton.me/drive.

Digitising the Past: Digital tools to support rock art research and dissemination

Ashely Green[1,2], Christian Horn[1,2], and Rich Potter[1,2]

[1] Department of Historical Studies, University of Gothenburg, Renströmsgatan 6, 41255 Gothenburg, Sweden
[2] Svenskt Hällristningsforskningsarkiv, University of Gothenburg, Renströmsgatan 6, 41255 Gothenburg, Sweden

Abstract

As digital applications in cultural heritage and rock art research continue to grow, resources from the Svenskt Hällristningsforskningsarkiv (SHFA) support such work by providing access to valuable data and tools. The SHFA archive includes data from the 17th century onwards and continually digitises the most recent documentation work. In collaboration with the Gothenburg Research Infrastructure for Digital Humanities we have created a platform to share this data with researchers and the public. The SHFA also provides tools and additional resources to visualise and enhance 3D data with Topography Visualisation Toolbox (TVT, https://tvt.dh.gu.se/). This empowers amateurs and professionals working with 3D recordings to improve their results. These visualisations can be used in deep learning workflows which drives the development of AI approaches in archaeology. In this paper, we provide an overview of SHFA's resources, their use in and beyond rock art research, their role in data dissemination, and future developments.

Keywords

Digital archaeology, data dissemination, visualisation, research infrastructure

1. Introduction

Digital approaches to rock art (including petroglyphs and paintings) documentation and analysis have been increasingly used over the last 15 years. From documentation to dissemination, digital methods have continued to improve rock art research and further our understanding of how rock art was created, the people who created it, its role in past societies, and how it can be viewed in modern times [1–3]. Digital methods largely encompass digitisation of traditional rock art documentation, 3D recordings (e.g., photogrammetry and laser scanning), statistical and geostatistical analysis, machine and deep learning, and dissemination platforms. Digital documentation, whether using Reflective Transformation Imaging (RTI), Structure from Motion (SfM) photogrammetry, or laser scanning, allows for additional analysis and extraction of carving patterns and motifs [4–6]. The use of geostatistical and GIS methods allows for both regional and large-scale analysis of rock art distribution and its role or positioning in the landscape [7–9]. Machine learning and deep learning approaches have been used recently in rock art research for both the identification and classification of motifs [10–13]. All of these methods and approaches culminate in the digital dissemination of rock art documentation and associated interpretations and metadata, as rock art requires explanation and interpretation. Digital tools, such as augmented reality (AR) [14], virtual reality (VR), and web-based platforms [15] provide data and access for researchers and the public.

Since 2007, the Svenskt Hällristningsforskningsarkiv (SHFA; Swedish Rock Art Research Archives) has led the development and implementation of digital methods and tools in rock art research. The SHFA has been a research infrastructure at the University of Gothenburg since 2017. It is responsible for documenting, archiving, and disseminating international rock art. Through collaboration and participation in research projects, the SHFA have developed tools, methodologies, and platforms to aid rock art researchers.

This paper presents an overview of the methods and tools from the SHFA and their current applications in rock art research, with a forward look on digital developments and applications beyond rock art research.

Huminfra Conference 2025, Stockholm, 12-13 November 2025.

ashely.green@gu.se (A. Green); christian.horn@gu.se (C. Horn); richard.potter@gu.se (R. Potter)

2. Tools

2.1. TVT

Topography Visualisation Toolbox (TVT; https://tvt.dh.gu.se/) is a set of tools that were originally developed to highlight the shallow carvings common in Bronze Age rock art in West Sweden [16]. However, it has continued to have broader applications within the cultural heritage sector. The tool works on similar principles to local relief modelling and difference maps for LiDAR data, but it provides additional features to generate the best possible visualisations of 3D documentation of planar surfaces (see **Figure 2**). It was developed for easy batch processing and user friendliness. This includes the base knowledge needed for both professionals and amateurs as not everyone is accustomed to programming or specialist software such as GIS software [17]. While it would be beneficial, not all archaeology departments employ technicians with coding experience to produce advanced visualisations. For researchers that simply want to produce visualisations, TVT provides an easy-to-use solution with a user-friendly interface, shown in **Figure 1**, to interact with the tools. The app is provided for Windows and MacOS, and the code is open source [18].

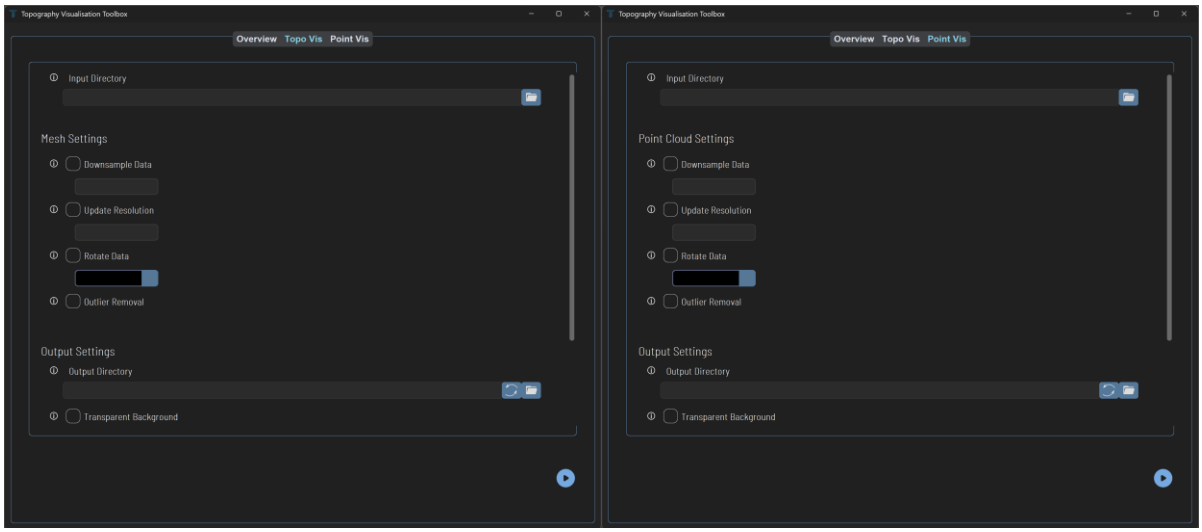


Figure 1: TVT app and tool GUIs for Windows.

At the core of TVT are two visualisation tools called Topo Vis and Point Vis. Topo Vis visualises 3D models, while Point Vis takes point data. Topo Vis includes additional processing steps to transform the mesh geometry into point data. Both use a difference map approach with additional contrast enhancements to highlight small changes in topography, for example carvings that are millimeters deep. Subtle traces like carvings are often obscured by larger changes in the local topography of the rock art panels. Difference maps help to remove larger, overpowering features, and thus, improve the visibility of the carvings. Users can select a range of optional settings, for example, when their data are scaled to metres rather than millimetres, as often occurs in photogrammetry 3D models, then they can use the *Update Resolution* setting to improve the visualisations. The 3D data (meshes or point clouds) must be stored in an input folder which allows for easy batch processing. The tools generate 16 visualisations and a summary of the data and processing settings which are saved automatically. Input and output folder locations can be easily adjusted in the app. The output files are shown in **Figure 2** and they include:

- depth map, texture map, normal map, derivative map
- topographic maps at two scales in greyscale and colour
- topographic maps with contrast enhancement using scikit-image's Contrast Limited Adaptive Histogram Equalization (CLAHE) function at two scales in greyscale and colour
- topographic maps which are a blend between the topographic map and enhanced topographic map at two scales in greyscale and colour

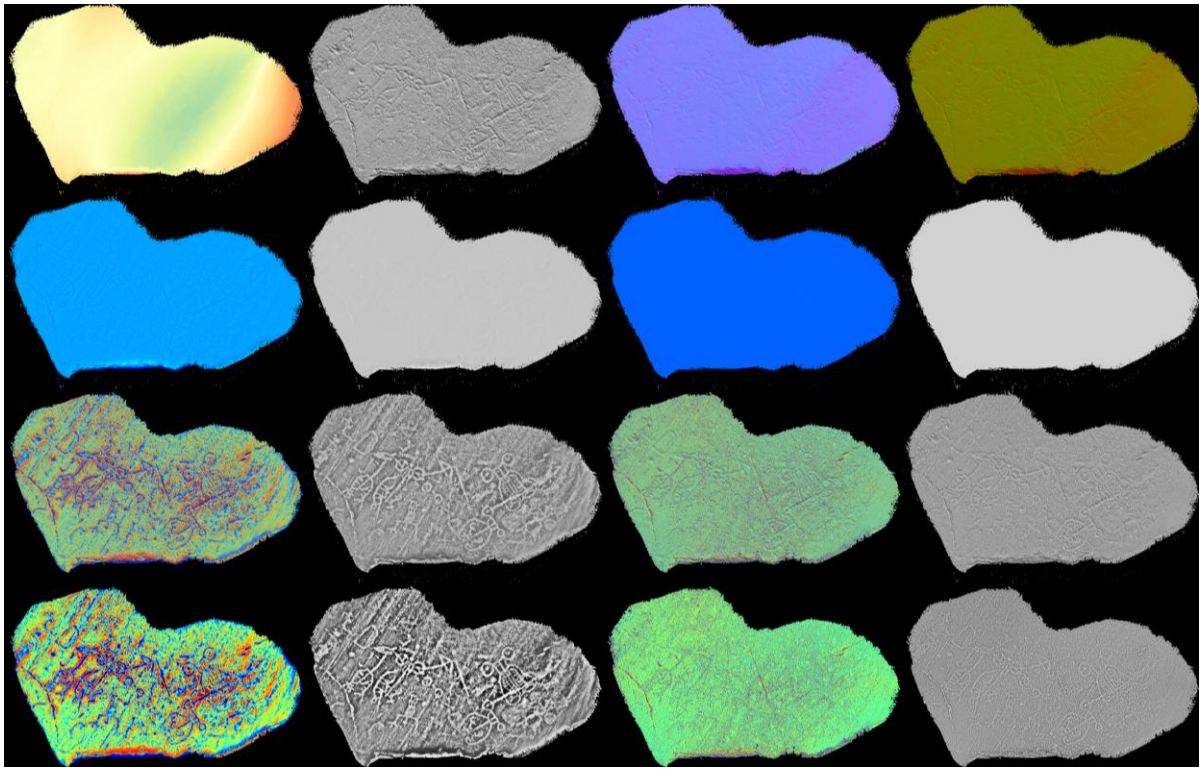


Figure 2: Example of TVT visualisations of a laser scan of the rock art panel Brastad 18:1. These visualisations demonstrate the potential uses of TVT to highlight shallow, small, and eroded carvings which are difficult to see in the original model. Original model by Ellen Meijer, visualisations by Ashely Green.

Depending on the data and the documented surface some of these visualisations will be more useful than others. The user will have to decide which visualisation(s) shows features in their data best. Processing the same data using the same settings in the same version of TVT will produce consistent results.

2.2. Tools for Deep Learning

Artificial intelligence (AI) approaches are growing in archaeology [19]. Rock art studies and other fields make use of geospatial data for these approaches, as such it is important to provide methods to generate training data in a streamlined workflow and subsequently test and understand the limits and challenges of these data in AI approaches. A toolbox for ArcGIS GIS software was developed to improve the workflow for exporting annotated data in YOLO OBB format [20] directly from the software. RockArtAITools [21] includes a script tool which tiles images and exports the clipped bounding box annotations. The tool allows users to draw their annotations directly in ArcGIS or import vectors from external sources, such as a Heritage Environment Record or national heritage board. The minimum bounding rectangle of the vector annotations are used in processing, so users can provide either rectangular bounding boxes or defined polygons in the tool. This tool builds on the existing annotation export tools provided by Esri, and more functionalities will be added for further AI applications.

2.3. SHFA Website and Database

The SHFA first launched a web platform around 2010. The updated SHFA database and website were developed in collaboration with the Gothenburg Research Infrastructure in Digital Humanities (GRIDH) in 2022 and officially launched in 2023. To date, the publicly available data includes nearly 27000 images and over 100 3D models from Sweden, Norway, Denmark, Spain, and Italy.

Development of the updated website was planned for four modules, with the first three completed as of 2024 and the final module available in late 2025.

The SHFA web resource is comprised of a PostgreSQL/PostGIS relational database in the Django framework and a website. REST APIs are used to retrieve data for the frontend. All images and 3D models in the database are related to a site, and associated 3D models and visualisations are grouped using a common identifier. Metadata stored for images includes:

- Image identifiers (numeric id and uuid)
- Location of the higher resolution IIIF display image
- Site identifiers and the site location information (e.g., coordinates, municipality, country)
- Group of panels or region an image belongs to
- Collection of images based on a common institution, region, or creator
- Original creator(s) and their affiliation(s)
- Year the image was taken or created
- Image type (e.g., photo, 3D visualisation, orthophoto, night photo) and subtype for visualisations
- Keywords to describe the motifs and image content, their associated terms in the Getty Art & Architecture Thesaurus controlled vocabulary, and whether the keyword describes a figurative motif
- Dating tags to describe archaeologists' interpretations of the motifs
- Visualisation group identifier

In addition to the metadata recorded for images, for 3D models we also record:

- Method
- Camera specifications, including focal length and crop factor, for SfM models
- Date and weather conditions for fieldwork
- Model dimensions, vertices, faces, and number of photos (for SfM models)
- Geology of the panel

The main website [22] uses the Vue3 framework, while the 3D model viewer [23] is separate from the main website to preserve compatibility with the 3DHOP library [24]. All resources are provided in Swedish and English. The main website (https://shfa.dh.gu.se/) is responsive and uses the split.js library to display three panels of data which increase in detail from left to right, as in **Figure 3**. The first panel contains the three search options – a free-text search, and advanced search, and a geographic/map search using OpenLayers [25]. The middle panel contains the search results. Once one of the search results is selected, the third panel is displayed.

The third panel contains the image metadata and, if available, a description of the site from the national heritage board or similar. Many fields in the metadata section are also clickable to trigger a new free-text search. All images that have an associated 3D model are indicated with a '3D' icon on the thumbnail in the search results gallery and a link to the model viewer is provided in the image metadata panel.

The 3D models are displayed in GRIDH's Multimodal Viewer [23], which uses open source libraries such as 3DHOP [24], Openseadragon [26], and OpenLime [27]. The model's metadata is displayed alongside the interactive 3DHOP and IIIF viewer, as in **Figure 4**. Users can change the lighting, navigate the model, measure the model, and, where available, turn the texture on/off.

Users can share a link to specific position in the 3D model viewer, for example to allow colleagues to easily discuss the same motif. Users can also share links to individual images and for a free-text search. Both the image and 3D model metadata sections include a suggested citation to allow users to easily align with the CC-BY copyright on all content. Images are downloadable and retain authorship, date, and id information in the filename.

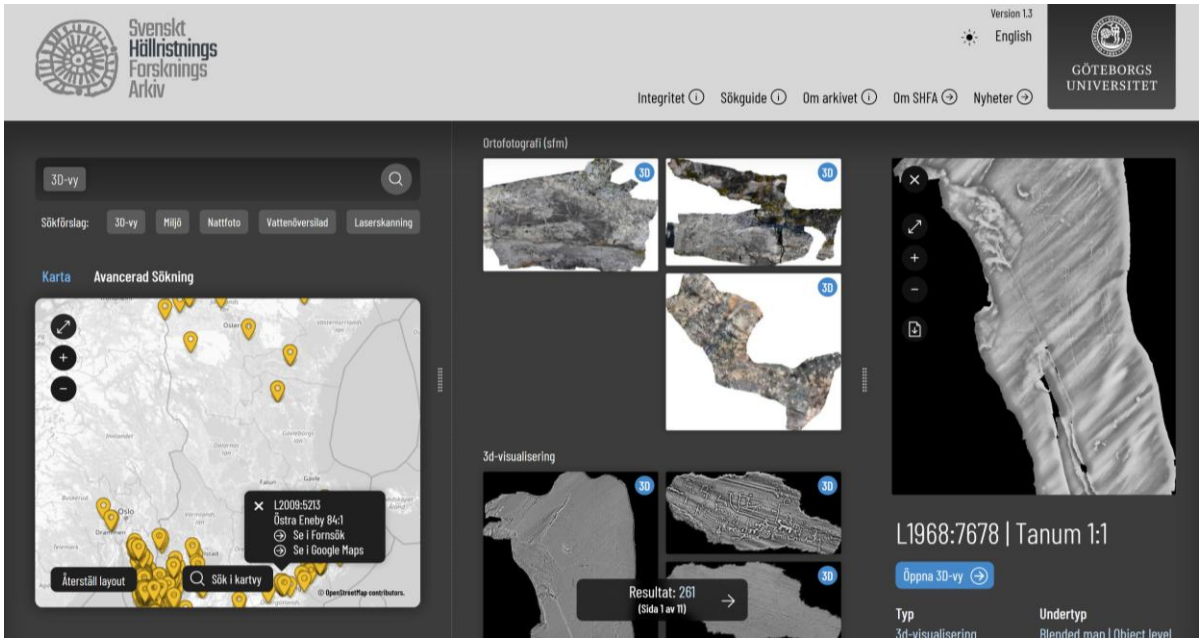


Figure 3: The SHFA website displaying the three-panel layout, highlighting the free-text and map search options, availability of 3D models, and the Openseadragon [26] IIIF image viewer.

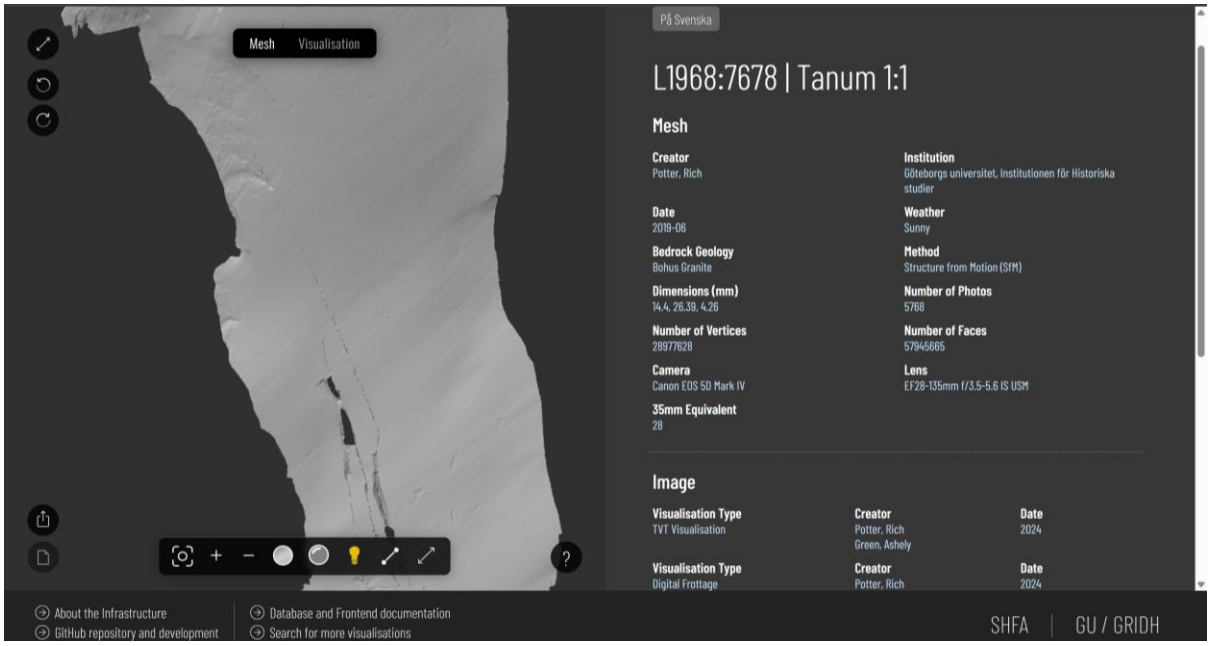


Figure 4: An example of the 3D model viewer and associated metadata.

3. Research & Impact

The SHFA and the tools it developed continue to support high-quality rock art research and other studies in which rock art serves as data. TVT enables researchers to produce visualisations of their 3D data quickly. Even considering the niche nature of TVT, it has had 172 downloads since 2024. TVT has been used by several other research groups with outputs published since 2022. The tools have been used successfully in research on Swedish, Norwegian, and US rock art, Iberian megaliths, Mesolithic and Early Neolithic material culture, Bronze Age cairns, for example. In these case studies, TVT not only provided material for research, but also the images to communicate the results to the readership. The visualisations offer a low-cost, readable format for displaying the data in publications as storage space is still a considerable cost factor and few publishers allow for embedding the full high-resolution 3D models in publications.

SHFA's main website aims to support research and dissemination on a broader scale than a tool aimed at the narrower segment of people engaging in 3D documentation. Within this scope, the launch of the new website can be seen as a huge success when considering what has been achieved since then. Beyond traditional rock art research, SHFA has supported such socially relevant research as difficult heritage studies, pedagogy, and library and information science among others [28–30]. It also has a significant role in aiding higher education by supporting theses on BA, MA, and PhD level.

Matomo is also used for analytics on the SHFA website. From this we can note the number of visits, interactions with the site, approximate location of visitors, and keyword search rates. As of September 2025, the new website has had over 30000 visits originating from 105 distinct countries. Visitors performed over 43000 searches. The top five search terms were in our suggested searches (skepp (ship), nattfoto (night photo), människofigur (human figure), djur (animal), hällristningsmiljö (rock art environment)), so we will implement randomized suggestions in 2025 to improve the user experience when revisiting the site. The statistics demonstrate consistent use of the SHFA archive and a growing use internationally.

Another measurement of the impact of the SHFA website and its tools is publications making use of these opportunities. For this, it should be kept in mind that the CC license under which we operate only demands that the publications cite the author of the images or documentations that are published. While we encourage authors to acknowledge SHFA in their publications, we depend on them doing so voluntarily and in a manner that can be tracked by search engines. Moreover, not all publications can be tracked. In total, at least 24 publications name TVT as a tool that was used since its first publication under its original name ratopoviz (rock art topography visualization) in 2022. This is a great achievement considering that TVT is a specialist tool without a large advertising footprint. SHFA has been mentioned since the launch of the new website in at least 80 publications. There is an increase of 50% from 2023 (22) to 2024 (33). While we cannot and should not expect a consistent increase in publications year-on-year, so far in 2025 we were able to identify 25 publications with SHFA material. These publications include at least 28 articles in scientific journals, 21 chapters in edited volumes, 16 theses, six monographs, five edited volumes, and four other publications such as fieldwork reports.

4. Discussion and Conclusion

The tools and platforms provided by the SHFA emphasise user-friendliness. Thereby lowering entry bars to access, for example, powerful visualisations for 3D data, deep learning in a GIS environment, or data. This benefits and furthers rock art research and the wider cultural heritage sector. By offering open-source tools and images under a less restrictive CC-BY license we aim to continue supporting the development of digital methods and workflows for rock art research. Further development of the SHFA website will focus on data summary tools and integration of external data, such as palaeoshoreline models from Sveriges geologiska undersökning (SGU) and site descriptions for international site (e.g., from Riksantikvaren for Norwegian panels). We will continue to share our data with Riksantikvarieämbetet and ARIADNEplus. These features will enhance the ability for researchers to use data hosted by SHFA in their projects and combine it with other datasets.

Acknowledgements

This work was funded by Riksbankens Jubileumsfond (grant no. IN18-0557:1, M21-0018) and the Swedish Research Council (grant no. 2020-01097, 2020-03817). We extend our thanks to all individuals working at the SHFA for supporting the progress that we have made. We also thank Jonathan Westin, Tristan Bridge, Aram Karimi, and Siska Humlesjö at GRIDH for their work in developing the SHFA database and website.

References

[1] M. Carrero-Pazos, R. Döhl, J.J. van Rensburg, P. Medici, A. Vázquez-Martínez, Rock Art Research in the Digital Era: Case Studies from the 20th International Rock Art Congress IFRAO 2018, Valcamonica (Italy), 2022. https://doi.org/10.30861/9781407360119.

[2] A. Green, C. Horn, Svenskt HällristningsForskningsArkiv Launches New Website, Curr. Swed. Archaeol. 32 (2024) 236–239. https://doi.org/10.37718/CSA.2024.17.
[3] C. Horn, M. Peternell, J. Ling, A. Green, R. Potter, Rock Art in Three Dimensions: Comments on the Use and Possibilities of 3D Rock Art Documentation, in: M. Hostettler, A. Buhlke, C. Drummer, L. Emmenegger, J. Reich, C. Stäheli (Eds.), 3 Dimens. Digit. Archaeol. State---Art Data Manag. Curr. Chall. Archaeol. 3D-Doc., Springer International Publishing, Cham, 2024: pp. 87–108. https://doi.org/10.1007/978-3-031-53032-6_6.
[4] M. Díaz-Guardamino, Rock Art Technology, Digital Imaging and Experimental Archaeology: Recent Research on Iberian Late Bronze Age Warrior Stelae, Complutum 34 (2023) 145–162. https://doi.org/10.5209/cmpl.85238.
[5] R. Potter, R. Rönnlund, J. Wallensten, An evaluation of Substance Painter and Mari as visualisation methods using the Piraeus Lion and its runic inscriptions as a case study, Herit. Sci. 11 (2023) 226. https://doi.org/10.1186/s40494-023-01071-7.
[6] M. Carrero-Pazos, B. Vilas-Estévez, A. Vázquez-Martínez, Digital imaging techniques for recording and analysing prehistoric rock art panels in Galicia (NW Iberia), Digit. Appl. Archaeol. Cult. Herit. 8 (2018) 35–45. https://doi.org/10.1016/j.daach.2017.11.003.
[7] T. Barnett, J. Valdez-Tullett, L.M. Bjerketvedt, F. Alexander-Reid, M. Hoole, S. Jeffrey, G. Robin, A Multiscalar Methodology for Holistic Analysis of Prehistoric Rock Carvings in Scotland, Herit. Sci. 12 (2024) 86. https://doi.org/10.1186/s40494-024-01183-8.
[8] J.L. Schaefer, A comparison of rock art and bluff shelter spatial distributions in the eastern Arkansas Ozarks, Southeast. Archaeol. 41 (2022) 1–15. https://doi.org/10.1080/0734578X.2021.2017636.
[9] M.L. Wienhold, D.W. Robinson, GIS in Rock Art Studies, in: B. David, I.J. McNiven (Eds.), Oxf. Handb. Archaeol. Anthropol. Rock Art, Oxford University Press, 2019: p. 0. https://doi.org/10.1093/oxfordhb/9780190607357.013.12.
[10] J. Kowlessar, J. Keal, D. Wesley, I. Moffat, D. Lawrence, A. Weson, A. Nayinggul, Reconstructing rock art chronology with transfer learning: A case study from Arnhem Land, Australia, Aust. Archaeol. 87 (2021) 115–126. https://doi.org/10.1080/03122417.2021.1895481.
[11] A. Jalandoni, Y. Zhang, N.A. Zaidi, On the use of Machine Learning methods in rock art research with application to automatic painted rock art identification, J. Archaeol. Sci. 144 (2022) 105629. https://doi.org/10.1016/j.jas.2022.105629.
[12] C. Horn, O. Ivarsson, C. Lindhé, R. Potter, A. Green, J. Ling, Artificial Intelligence, 3D Documentation, and Rock Art—Approaching and Reflecting on the Automation of Identification and Classification of Rock Art Images, J. Archaeol. Method Theory 29 (2022) 188–213. https://doi.org/10.1007/s10816-021-09518-6.
[13] C. Horn, A. Green, V.W. Skärström, C. Lindhé, M. Peternell, J. Ling, A Boat Is a Boat Is a Boat…Unless It Is a Horse – Rethinking the Role of Typology, Open Archaeol. 8 (2022) 1218–1230. https://doi.org/10.1515/opar-2022-0277.
[14] J. Westin, A. Råmark, C. Horn, Augmenting the Stone: Rock Art and Augmented Reality in a Nordic Climate, Conserv. Manag. Archaeol. Sites 23 (2021) 258–271. https://doi.org/10.1080/13505033.2023.2232416.
[15] A. Green, T. Bridge, C. Horn, S. Humlesjö, A. Karimi, J. Ling, J. Westin, Accessing centuries of documentation - Resources to improve access to Swedish rock art documentation and metadata, in: Proc. Huminfra Conf. HiC 2024, Linköping University Electronic Press, Linköping, 2024: pp. 154–160. https://doi.org/10.3384/ecp205021.
[16] A. Green, C. Horn, R. Potter, Topography Visualisation Toolbox: Project and Software Summary, (2025). https://tvt.dh.gu.se/ (accessed August 20, 2025).
[17] R. Potter, D. Pitman, L. Shaw, C. Horn, Everyone Has to Start Somewhere: Democratisation of Digital Documentation and Visualisation in 3D, Open Archaeol. 11 (2025) 20250054. https://doi.org/10.1515/opar-2025-0054.
[18] A. Green, O. Ivarsson, R. Potter, C. Horn, Topography Visualisation Toolbox (TVT), (2025). https://doi.org/10.5281/zenodo.15479454.
[19] S.H. Bickler, Machine Learning Arrives in Archaeology, Adv. Archaeol. Pract. 9 (2021) 186–191. https://doi.org/10.1017/aap.2021.6.
[20] G. Jocher, J. Qiu, Ultralytics YOLO11, (2024). https://github.com/ultralytics/ultralytics.

[21] A. Green, R. Potter, C. Horn, RockArtAITools, (2025). https://arcg.is/191Gjn1 (accessed April 30, 2025).
[22] J. Westin, T. Bridge, A. Karimi, A. Green, S. Humlesjö, SHFA Frontend + Backend, (2025). https://github.com/gu-gridh/shfa-frontend (accessed October 9, 2025).
[23] J. Westin, T. Bridge, A. Green, J. Beck, gu-gridh/multimodal-viewer, (2025). https://github.com/gu-gridh/multimodal-viewer (accessed August 20, 2025).
[24] M. Potenziani, M. Callieri, M. Dellepiane, M. Corsini, F. Ponchio, R. Scopigno, 3DHOP: 3D Heritage Online Presenter, Comput. Graph. 52 (2015) 129–141. https://doi.org/10.1016/j.cag.2015.07.001.
[25] openlayers/openlayers, (2025). https://github.com/openlayers/openlayers (accessed August 20, 2025).
[26] I. Gilman, A. Kishore, C. Thatcher, M. Salsbery, A. Vandecreme, T. Pearce, OpenSeadragon, (2024). https://github.com/openseadragon/openseadragon (accessed August 20, 2025).
[27] cnr-isti-vclab/openlime, (2025). https://github.com/cnr-isti-vclab/openlime (accessed August 20, 2025).
[28] G. Andersson, E. Bylin, Vetenskaplig publicering inom humaniora : En jämförelse av två fördelningsmodeller, 2024. https://urn.kb.se/resolve?urn=urn:nbn:se:hb:diva-32493 (accessed September 4, 2025).
[29] M. Legnér, Difficult Heritage or Objects of Science?: The Material Legacy of Nazi German Rock Art Research in Sweden, in: P.B. Larsen, M. Křížová (Eds.), Eur. Univ. Legacies Probl. Herit. Contemp. Pract., Edinburgh University Press, Erscheinungsort nicht ermittelbar, 2025: pp. 197–220.
[30] L. Almqvist Nielsen, Prehistoric history in Swedish primary school education: pupils' expression of empathy after visiting a cultural heritage site, Educ. 3-13 53 (2025) 378–392. https://doi.org/10.1080/03004279.2023.2191631.

AI Pedagogy and New Methods for Humanities Scholars: A Reflective Case Study

Chris Haffenden[1], Justyna Sikora[2,*]

[1]KBLab, Kungliga biblioteket, Karlavägen 100, 115 26 Stockholm, Sweden
[2]KBLab, as above

Abstract

This paper presents a reflective case study on teaching AI methods to humanities scholars through the example of a workshop developed at KBLab, the AI and digital research lab at the National Library of Sweden. Conducted within the framework of the Swedish national research infrastructure Huminfra, the workshop introduced participants from heritage organisations and academia to multimodal topic modelling using CLIP-Topic and Google Colab. Drawing on this experience, we discuss pedagogical strategies for bridging the divide between large-scale computational analysis and the interpretive traditions of the humanities. We reflect on design choices such as how much code to expose, how to balance explanation and demonstration, and how to make domain-specific applications intelligible to non-technical users. The case illustrates how script-based, open formats can foster AI literacy and critical engagement, while highlighting the limits of one-off workshops and the need for sustained, cross-disciplinary collaboration.

Keywords

AI pedagogy, AI literacy, multimodal topic modelling, digital humanities, cultural heritage collections,

1. Introduction

How can we bridge the divide between increasingly large-scale computational approaches to GLAM collections and the more qualitatively inclined perspectives of humanities scholars [1], while also raising awareness of the possibilities and limitations of AI-based search systems now entering the research landscape? What pedagogical strategies are available to those of us who work hands-on with digital research infrastructures and research services, seeking to connect these methodological worlds?

This paper draws on our experience at KBLab, the AI and digital research lab at the National Library of Sweden [2], and our outreach initiatives within the Swedish national research infrastructure Huminfra. We discuss the design and delivery of a new workshop developed to illustrate the potential of multimodal topic modelling for both heritage organisations with large collections of unlabelled images, and researchers working with visual culture.

The workshop builds on our earlier work with using **CLIP** (Contrastive Language–Image Pre-training) to enhance the searchability of the library's postcard holdings [3] and with creating a user-friendly, script-based workshop for **BERTopic** [4]. In this paper, we reflect on our design choices in using Google Colab to demonstrate how image collections can be clustered according to topic—for instance, how much code to include and how many steps can realistically be explained. We also discuss the affordances and limitations of web-based workshops as a mode of methodological outreach.

By focusing on this concrete scenario, we highlight how AI literacy efforts in the humanities can be rooted in domain-specific needs and interpretive questions. We argue that effective outreach requires not only demystifying technical tools but also fostering dialogue around how these tools intersect with established scholarly practices. The case study illustrates the value of flexible, script-based formats that empower researchers to engage critically with emerging methods without requiring full technical

Huminfra Conference 2025, Stockholm, 12–13 November 2025.
*Corresponding author.
chris.haffenden@kb.se (C. Haffenden); justyna.sikora@kb.se (J. Sikora)
0000-0002-5561-5163 (C. Haffenden)

fluency. At the same time, we remain cautious about the transformative potential of one-off workshops, emphasising instead the importance of longer-term, transdisciplinary forms of methodological collaboration.

2. CLIP-Topic workshop: translating infrastructure into pedagogy

Before describing the workshop itself, we briefly place it within the broader context of digital research infrastructures. Initiatives such as KBLab and Huminfra increasingly act not only as technical enablers but also as pedagogical intermediaries: they are often where humanities scholars encounter AI methods in depth for the first time, bringing questions of usability, literacy and interpretation to the fore. In this sense, infrastructures have become *pressure points* for digital pedagogy—sites where the demands of large-scale computation meet the interpretive expectations of the humanities. We conceived of our CLIP-Topic workshop within this intersection, using the lab's resources and expertise to translate an advanced AI method into an accessible, conceptually meaningful learning experience.

The AI method demonstrated in the workshop rests on a simple but powerful idea: *vector-based search*, which allows computers to measure similarity not through exact matches, but through proximity in meaning or appearance. In the practical setting of multi-modal heritage collections, this means that both words and images can be converted into numerical representations—known as *embeddings*—and compared within a shared vector space. Items that are close to one another in this space are interpreted as being semantically or visually related. This principle underpins the recommendation engines, image searches, and content-suggestion systems that most of us interact with daily— when a smartphone groups photos by theme, or when a streaming platform suggests films "you might also like". The same technique enables us to cluster and explore large-scale heritage collections by their contents rather than their (often scant) metadata, offering new entry points into archives that have long remained opaque to search [3].

2.1. Workshop aims

Our aim with the workshop in broad terms was to showcase how the same AI techniques we encounter in everyday online contexts can be harnessed to improve the accessibility and research potential of image collections. Beyond introducing the conceptual foundations of such techniques, we also wanted to convey the practical mechanics of working programmatically with "collections as data"—that is, through a hands-on immersion in running code.

The target audience consisted of two groups, partially overlapping but principally divergent: heritage professionals with responsibility for digital collections management, and academic scholars in the humanities and social sciences working with digital images. The learning objectives for these groups differed slightly. In both cases, we wanted to demonstrate a sense of the possibilities of using vector search to cluster and organize a large volume of documents—in this case, images—according to their thematic contents rather than preexisting (and often limiting) categories [5]. More specifically, for heritage professionals we aimed to offer tangible inspiration for applying AI-based search to their own collections, and for researchers, to provide a concrete sense of how such techniques can serve as a starting point for exploring and analyzing visual research data.

2.2. Designing the workshop

To enable participation on a national scale, we chose to make the workshop web-based, ensuring that heritage professionals and researchers outside the Stockholm area could take part. Conceived as a pedagogical activity delivered via Zoom, the workshop followed a three-part structure, which we outline below.

The first part consisted of a *conceptual introduction*—a twenty-minute presentation that familiarised participants with two key components of the workshop. First, we outlined the particular challenges facing heritage institutions with large collections of unlabelled images, to which CLIP-Topic offers a potential

solution. The aim here was to anchor what might otherwise seem a highly technical method within a concrete practical context, underlining that such tools are valuable insofar as they help us address the real challenges of working with digital culture. Second, we provided a brief overview of the conceptual architecture of the method, beginning with an explanation of OpenAI's CLIP model [6] and followed by a sketch of the logic behind topic modelling [7]. The goal was to give participants a basic understanding of multimodal topic modelling without confusing or alienating non-technical attendees with excessive detail; accessibility and broad brushstrokes were deliberately prioritised over depth. (For those interested in exploring the method in more depth, we included a slide with links to further reading.)

The second and main part of the workshop consisted of a *live demonstration*: a guided walkthrough of a Google Colab notebook we created to show how CLIP-Topic can cluster and visualise images by topic [8]. We chose Colab as the pedagogical platform because it allows participants to run pre-written Python notebooks directly in the browser, while also providing free access to the GPUs needed to fit the models. This approach removes typical barriers of software installation and environment setup, while keeping the process transparent—every cell of code can be read, edited and executed. We had previously used Colab for workshops on text-based topic modelling with BERTopic [4], where participants particularly appreciated being able to revisit and rerun the notebooks at their own pace. The same logic guided this design.

To ensure it could be published as an open resource—something participants could return to and others discover and use on their own terms—we based the workshop on openly available Swedish heritage data. The notebook uses a dataset of 1,508 images downloaded from DigitaltMuseum. We selected these as situated examples: by choosing heritage images likely to be familiar to participants in both style and sentiment, we hoped to anchor the more technical discussion of AI models in the professional realities of those taking part. In addition to emphasizing that the notebook can be revisited for future learning, we also highlighted its adaptability—users are encouraged to insert their own data into the script and run the analysis on their specific collections.

The notebook itself was organised to alternate between short explanatory sections and executable code cells, allowing participants to see each step of the process before running it. This structure was intended to make the workflow intelligible even to those without prior programming experience, while still offering enough detail for more technically confident users to experiment further.

The third and final part centered on *applications and questions*. After completing the Colab walkthrough, we shifted focus from the mechanics of the method to what it can reveal in practice. Using examples drawn from ongoing experiments with image search and description at both KBLab and Stanford University Library, we demonstrated how vector-based techniques can surface new thematic connections within heritage collections and support visual exploration at scale [9]. This segment also served as an open discussion space: participants were invited to reflect on how such approaches might translate to their own institutional or research contexts, and to raise questions about, for instance, interpretability. In pedagogical terms, the aim was to re-connect the technical and conceptual threads of the workshop—moving from doing to thinking, and from method to meaning—thereby reinforcing AI literacy as both a practical and interpretive competence.

This three-part format—combining conceptual framing, guided demonstration and open discussion—was designed to balance accessibility and depth. It aimed to help participants connect technical operations with interpretive reflection and imagine how AI methods might be meaningfully adapted to their own disciplinary and institutional settings.

3. Pedagogical reflections: Successes and Challenges

When we first ran the workshop, over 60 people had registered in advance and more than 40 attended the two-hour online session. The fact that participants came from such a wide range of institutions—from major national museums to small local archives, and from universities stretching from Florence and Cork to Luleå—attests both to the topicality of the subject and the benefits of circulating the invitation through diverse scholarly and professional networks. In what follows, we offer some reflections on what worked

well with the workshop as a pedagogical event, and what proved more challenging.

3.1. What worked

Transparency and legibility of code. In an era of chat-based LLMs offering researchers immediate results but little accountability for how those results are produced, one of the key advantages of a step-by-step, script-based workshop was the *sense of process* it provided. By including every stage—from loading data and producing embeddings, to fitting the model and visualising the results—our CLIP-Topic notebook helped participants grasp the logic of the entire workflow. Including each line of executable code also conveyed that this was about the operation of a legible method, rather than the "silver-bullet" magic of AI. Through combining explanatory comments with code cells, we could walk participants through three modes of topic modelling images—using only textual descriptions, only the images themselves, and a combination of the two—accompanied by a running commentary on the distinct effects and merits of each approach.

Advanced access as icebreaker. From previous experience, we have learned the value of sharing the notebook a few days before the workshop. This gives participants the opportunity to explore the code and familiarise themselves with the method in advance. Such early access can be particularly beneficial for non-technical users who might otherwise feel overwhelmed by lines of Python. The advantage is that participants can then focus on the live explanations during the session, rather than feeling stressed about running code cells.

This approach aligns with our goal of keeping the event accessible to all, regardless of prior programming experience: the technical threshold is deliberately low. It also underlines the notebook's sustainability as a resource that participants can return to and adapt after the workshop.

Situated examples and relevance. Following the script demonstration with examples of domain-specific applications—vector-based image search and VLM-generated image descriptions—helped situate the method within the realities of heritage collections. Showing how we had already integrated CLIP into an image search demo at the National Library provided a concrete reminder that such approaches are feasible in practice, rather than speculative.

It also prompted relevant questions from heritage professionals during the final discussion. That said, such a "shopwindow" presentation of applications carries a risk of oversimplifying the scale of development work involved—a point we return to below.

3.2. What proved difficult

Lack of interactivity and real-time feedback. The principal drawback of offering the workshop to a large, technically diverse audience online was that participants tended to take part fairly anonymously, risking becoming passive listeners rather than active learners. The online setting creates a challenging pedagogical situation: talking through and explaining a series of code cells can at times feel like lecturing a brick wall. Given that the schedule was tightly packed, there were perhaps too few opportunities to check in with participants during the script to gauge whether they had questions or reflections.

We encouraged the use of the chat for questions, but this is not equivalent to a classroom experience where instructors can identify and respond to moments of confusion in real time. The lack of interaction was probably accentuated by the fact that participants may not always feel comfortable raising questions in such settings, particularly when the group spans a wide range of skill levels and institutional contexts.

Limits of single-session workshop. This challenge of limited interactivity is, to a significant extent, an effect of the format itself. There are strict limits to what can be achieved in a two-hour session—especially one that starts from scratch. The ambition to showcase both the conceptual and practical possibilities of the method necessarily reduced the time available for dialogue and experimentation.

In previous workshops, we have addressed this issue by dividing the material into two sessions, with a "homework" task in between: a first session devoted to theory, an individual exercise in between, and a second session for reflection and discussion. In future iterations, we may consider introducing a similar

structure for this workshop as well—though the trade-off is that it becomes harder to attract participants to commit to multiple dates.

Risk of appearing "plug and play". A final challenge concerns the risk of oversimplifying what such methods entail. The self-playing quality of a Colab notebook can give the impression that multimodal topic modelling is a "plug-and-play" solution: press run, and AI does the rest. In reality, deploying these methods within heritage organisations is rarely frictionless. Adapting them to local infrastructures and workflows requires sustained technical maintenance and coordination across organisational silos. In this sense, the workshop risked presenting AI as a normalised, readily available technology when, in practice, it remains dependent on scarce expertise, time, and institutional support.

A key pedagogical takeaway, therefore, was to demystify not only the models themselves but also the organisational conditions that shape their use in practice. We raised this point in response to a question, but the exchange underscored the importance of addressing it explicitly as a condition of possibility within the session itself.

Broader reflections. Taken together, these challenges point to a broader insight: effective AI pedagogy in the humanities depends as much on institutional capacity, infrastructural support, and collaboration as on individual technical skill or curiosity. It is less a matter of training humanities scholars to become AI experts, and more about fostering critical and interpretive literacy—while strengthening the conditions for sustained interaction between researchers, data scientists, and the institutions that mediate between them.

4. Conclusion and further work: towards AI literacy

Designing and running this type of introductory workshop involves a considerable degree of pedagogical risk. Participants arrive with widely varying levels of technical experience; even a minimal Colab notebook can cause cognitive overload for those new to programming; and the very convenience of the format can create an illusion of understanding without deeper comprehension of the underlying mechanics. Acknowledging these risks from the outset, we sought to balance transparency and "data realism" with accessibility and usability.

Our experience affirms a broader insight within the digital humanities learning community: workshops are powerful entry points but weak sustainers. They can ignite curiosity and provide conceptual orientation, yet they remain far removed from the pedagogical structures of sustained practice, repetition, and peer learning required to develop technical proficiency. Nor should we expect them to deliver miracles or shortcuts to understanding. If we instead regard workshops as gateways to longer processes of discovery and engagement, their value becomes clearer: they equip participants with a new vocabulary, methodological awareness, and sense of possibility.

In terms of further work, we are planning several adjustments for the next iteration of the workshop. First, we intend to experiment with a smaller, more interactive format—either as a limited online group with greater opportunities for exchange, or as an in-person teaching event, for instance within a master's programme in Digital Humanities. Second, we will explore adding short reflective checkpoints within the Colab notebook to prompt active engagement during the session. Building on this, we are also considering a follow-up meeting to provide more time for discussion and cross-participant exchange. Finally, we are considering introducing a post-workshop feedback form to gather more systematic insights into participants' experiences.

The broader task ahead lies in embedding AI literacy within the interpretive habits of the humanities. This means treating computational methods not as external aids but as evolving objects of inquiry in their own right—tools that both enable and transform the ways we encounter and understand cultural heritage collections. Workshops like this represent a modest but necessary step toward that goal—and toward a more reflective, literate engagement with AI in scholarly and heritage practice.

References

[1] L. Jaillant, K. Aske, Are Users of Digital Archives Ready for the AI Era? Obstacles to the Application of Computational Research Methods and New Opportunities, J. Comput. Cult. Herit. 16 (2024) 87:1–87:16. URL: https://dl.acm.org/doi/10.1145/3631125. doi:10.1145/3631125.

[2] L. Börjeson, C. Haffenden, M. Malmsten, F. Klingwall, E. Rende, R. Kurtz, F. Rekathati, H. Hägglöf, J. Sikora, Transfiguring the Library as Digital Research Infrastructure: Making KBLab at the National Library of Sweden, College & Research Libraries 85 (2024) 564. URL: https://crl.acrl.org/index.php/crl/article/view/26325. doi:10.5860/crl.85.4.564, number: 4.

[3] C. Haffenden, F. Rekathati, E. Rende, Unearthing forgotten images with the help of AI – The KBLab Blog, 2023. URL: https://kb-labb.github.io/posts/2023-10-20-unearthing-forgotten-images-with-the-help-of-ai/.

[4] F. Rekathati, BERTopic Workshop: Analyzing Swedish Parliamentary Motions, 2022. URL: https://colab.research.google.com/drive/10kB3wfoHSfZE48vEKmznIw-ff36uR8gs?usp=sharing.

[5] M. Malmsten, V. Lundborg, E. Fano, C. Haffenden, F. Klingwall, R. Kurtz, N. Lindström, F. Rekathati, L. Börjeson, 13 Without Heading? Automatic Creation of a Linked Subject System, in: E. Balnaves, L. Bultrini, A. Cox, R. Uzwyshyn (Eds.), New Horizons in Artificial Intelligence in Libraries, De Gruyter Saur, Berlin, Boston, 2025, pp. 179–198. URL: https://doi.org/10.1515/9783111336435-014. doi:doi:10.1515/9783111336435-014.

[6] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, I. Sutskever, Learning Transferable Visual Models From Natural Language Supervision, 2021. URL: https://arxiv.org/abs/2103.00020v1.

[7] D. M. Blei, Probabilistic topic models, Communications of the ACM 55 (2012) 77–84. URL: https://dl.acm.org/doi/10.1145/2133806.2133826. doi:10.1145/2133806.2133826.

[8] J. Sikora, AI for Image Collections: A Hands-On Workshop with CLIPtopic, 2024. URL: https://colab.research.google.com/drive/1CMJ0ko8kWA4Zjf_qkBA0AN_TwDHeRV8t?usp.

[9] P. Broadwell, L. King, Beyond "This Image May Contain...", 2025. URL: https://thegogglesdonothing.com/projects/cni2025/.

Documenting AI Use in Humanities Research

Isto Huvila[1]

[1] *Department of ALM, Uppsala University, Thunbergsvägen 3H, Uppsala, Sweden*

Abstract

This paper explores the critical need to document the use of Artificial Intelligence (AI) in humanities research. While AI offers efficiency and analytical power, its application raises concerns about transparency, bias, and reproducibility. Existing documentation frameworks often emphasise technical aspects, overlooking the human and contextual dimensions vital to humanities scholarship. Drawing on cross-disciplinary literature, the paper advocates for integrating paradata (process-related meta-information) to capture both technical and human facets of AI use. It proposes shifting the focus from speculative future needs to documenting the transformation AI is intended to achieve within specific research contexts. Practical strategies include combining automated tools with reflective documentation practices and providing clear explanations of the purpose and expected outcomes of AI use. The paper calls for infrastructural support and a rethinking of documentation sufficiency to enhance understanding, reuse, and accountability in humanities research.

Keywords

paradata, artificial intelligence (AI), documentation

1. Introduction

Adequate documentation of Artificial Intelligence (AI) systems is critical for the transparency and accountability of their use and understandability of their outcomes in humanities research. AI offers multiple benefits, including of being potentially less expensive and more efficient to use than competing techniques of processing data. In addition, they are argued to be less biased and capable of achieving higher levels of accuracy in various tasks [1, 2] many of which are pertinent to humanities scholarship. At the same time, there are, however, many well-known risks. Outcomes of AI systems can be unpredictable, difficult to understand and fallacious [3, 4, 2]. AI systems have also shown tendencies to discriminate, and to replicate and perpetuate historical societal biases [5]. In spite of the critical importance of explainability of AI and AI use [4], documenting AI use in humanities research to avoid production of biased and fallacious data and knowledge, so far the bulk of the emerging research on documenting both technical and social aspects of AI systems and their use has focused on other fields of scholarship (cf. e.g., [6, 7]).

The aim of this paper is to draw attention to the critical importance of documenting AI use and include such documentation both in terms of documents and the practice of documenting AI use in infrastructures that support humanities research. Drawing on a cross-disciplinary reading of the literature with a focus on insights relevant to AI use in humanities context, this paper calls attention to why documenting technical characteristics of AI systems is not enough and proposes measures to improve the usefulness of the documentation of the human aspects of AI use.

2. Documentation of research processes in the AI era

Recent literature has been increasingly emphasising the importance of not merely documenting research outputs but also the process of how research is conducted. Adequate process documentation in terms of diverse forms of paradata (understood broadly as process information

Huminfra Conference 2025, Stockholm, 12-13 November 2025.

isto.huvila@abm.uu.se (I. Huvila)

or a "category of things that can be appropriated as informative of processes and practices" [8]; for a more in-depth discussion of the concept, see [9]) is crucial to understanding and accountability of research practices, outputs and the reusability of research data. At the same time, knowing and eliciting paradata that is useful for sometimes hard to predict future uses and determining how much paradata is enough can be exceedingly difficult [10].

A critical part of paradata is sufficient documentation of the use of any technologies applied in the research process. This applies to, for example, measurement and other equipment used for data capturing but also software and instruments used in data analysis and producing research outputs [10]. Current rapidly advancing AI technologies present a new set of opportunities for paradata generation [11] but also challenges to documenting the technologies themselves when applied in research processes [12]. Much of the work relating to AI use in research has focused partly on debating what types of AI use are ethical and partly on how to properly acknowledge AI use. Much less attention has been directed on how and what to document about the AI in the loop [13, 7] and even fewer studies have addressed so far the issue of AI documentation in the context of humanities research (exceptions e.g. [14, 15]). Even if much remains to be done in developing both theoretical and practical understanding of documenting AI use in research processes, there is an emerging transdisciplinary body of work relating to AI documentation [7], a part of which can help to provide relevant insights for paradata work also in humanities scholarship.

The principal concern of AI regulation and documentation has been on the diverse uses of AI in contexts judged as particularly risky pertaining to health, safety and fundamental human rights [2]. A growing number of frameworks and guidelines provide general guidance and directives of what is expected to be documented relating to AI systems and their use. The existing guidelines have, however, criticised of having multiple shortcomings. Many of the frameworks are focused on technical documentation of tools rather than their use [3] and how they integrate in human processes [16]. Similarly, a dominant theme in earlier observations is how documentation and many documentation guidelines give precedence to accountability, managing risks and liabilities and conformity to rules and regulations (e.g., [3][17]) or technical integration of systems [18] rather than promoting reproducibility and understanding and reuse of the outcomes of AI use. Königstorfer [7] notes further that proposed documentation techniques are seldom evaluated by governance experts or auditors. In parallel to high technical specificity in part of the proposed techniques, the frameworks that focus on outlining general principles of ethical AI implementation risk to be too unspecific to provide actionable guidance [19][10][7]. Privacy, research ethics and lack of permission to share process information are also identified as major impediments especially in proprietary contexts and research with human subjects involving sensitive personal data [20][21].

The inclination of AI documentation techniques to focus on documenting technical artefacts and preserving easily documentable aspects of AI use parallels with observations on research data makers' similar concern when documenting data making, processing and use [22]. Their focus of interest differs from data reusers' emphasis of information that makes data understandable and reusable [22]. This parallels with concerns expressed of the usability and usefulness of software-focused AI documentation produced according to many of the currently proposed frameworks [16].

3. Approaches to the documentation of AI use

Regarding the aspects of AI and AI use that needs to be documented, unsurprisingly considering the critique of technology overemphasis, the literature underlines the need to document the technical design and functionality of AI systems, including the AI model [6], and its training data [16]. In addition, the documentation should include information on the application context of the system and its development process [16]. Königstorfer [7] has analysed AI documentation techniques and identified three main aspects of AI that are recurringly in the focus, including the documentation of training data, application domain and design decisions. Moreover, similarly to the literature on research documentation and sharing (e.g., [9][23]), the AI literature underlines the importance of the understandability of AI documentation and balancing the costs and benefits of producing it [16].

Techniques proposed for documenting training data include approaches based on providing summary statistics and visualisations of datasets. Tools exist for analysing diverse aspects of datasets, including bias and fairness, and facilitating the creation of relevant documentation [24, 25, 26]. Unstructured data is generally more difficult to document and often relies on descriptive metadata rather than summary data [7]. The techniques proposed for collecting concise summary documentation of the paradata on the use of AI, including various aspects of collecting and processing training data, application domain, design decisions, and deployment and use of AI are predominantly based on guidelines formulated as templates and checklists. Such frameworks include, for example, such frequently cited techniques as Model Cards [26] and FactSheets [13]. There are also tools like Jupyter Notebooks and workflow tracking systems that allow to diverse extents automatic documentation of computational processes [27][28].

In addition to AI specific documentation techniques, the majority of the methods applicable for general paradata documentation are also useful for documenting AI use. Liu and Huvila [11] provide examples of categories of such methods including methods descriptions in form of formal metadata, narrative descriptions, recording, logging, research plans and prospective workflows. As suggested by Juneström and Huvila [29], such approaches and multiple others can be used to document paradata prospectively before, during, and retrospectively, after AI use.

4. Challenges and ways forward

As with paradata in general, the key challenges with AI use specific paradata are how to elicit adequate documentation and how to determine what to preserve (cf. [10]). Both earlier paradata [30, 22, 31] and AI documentation research [7] point to that while user needs are a critical premise of what documentation is relevant, they are also a fleeting target. Expressed needs vary between different users and uses that are associated with diverging perspectives to what aspects of AI use require explanation and transparency. Humanities research is generally characterised by a greater diversity of research practices than paradigmatically more homogenous disciplines [32, 33]. This accentuates the challenges of documenting AI use. It is unlikely that a single documentation standard could capture all information relevant for all humanities research, or even for a single discipline or discipline-antagonistic "data culture" [21].

Much similarly to process documentation in general [10], undoubtedly the most critical challenge of documenting AI use especially in humanities research context is how to capture enough while adding as little to the effort of documentation as possible. When documenting AI use, it is important to try to strike a balance between focusing on keeping (relatively) easily collectable and preservable technical information directly from and in research infrastructure services and identifying, generating and keeping typically more eclectic information about the use of AI to an extent that is necessary for understanding the systems and its use. Keeping certain technical information, such as summary information of training data and design specifications of models and algorithms, on AI algorithms and uses on the basis of its easy collectability and preservability is sensible. Similarly to collection of paradata in general [11], a part of the documentation of AI systems can be automated [7]. In contrast to keeping such information, assuming that it would automatically be enough is, however, highly problematic [11]. The ease of keeping specific types of information also comes with an immanent risk of keeping too much that makes searching and preserving the information both too laborious and resource-demanding [34]. Instead of assuming that keeping existing information would solve the conundrum, a critical question to consider is the sufficiency of documentation. In cases when sufficiency can be to a reasonable extent specified, evaluated and audited, such measures should be considered. In contexts like the humanities research where a single measure of sufficiency is close to an oxymoron due to the diversity of practices and theoretical perspectives, an approach worth considering is to try flip the perspective. Instead of approaching sufficiency from imagined and sometimes unimaginable user perspectives, a more practical line of action ought to be to approach it from the premises of how a specific instance of AI use is conceived from the perspective of the AI use itself.

First, instead of trying to produce documentation for others and assuming that it provides expected type of transparency, a potentially useful complementary strategy is to try to produce a *brief description of what each piece of documentation is attempting to achieve and how*. For example, such documentation might contain a brief explanation of how a descriptive summary of training data and a narrative description of the principles of how an algorithm works are expected to allow an individual with basic skills in computational methods to understand the logic of how the documented system generates its outputs.

Second, rather than assuming that a meticulous step-by-step description of what was done could capture all relevant information, a potentially powerful complementary approach is to produce a *documentation of what a particular AI system was used for and what it was considered to have achieved in the context of its use*, that is, what was the purpose of AI use and how the outputs and outcomes of an AI system were considered to compare to its inputs. A pertinent critique of present process documentation practices is that it is not always clear what documentation is attempting to achieve [35]. The general contextuality of process documentation [36] increases the risk that acontextual procedural descriptions are especially unclear in this respect. In contrast, when the focus is put on explicating the transformation AI is used to achieve, the likelihood of being able to convey a more comprehensive understanding of the *use* of AI can be expected to be greater.

The crucial step in switching perspective from speculating on future user needs in cases when they are unknown to taking the AI use itself as a starting point is not to produce two new categories of documentation but rather to take a perspective that can be taken and utilising it as a starting point for producing meaningful documentation. This is also a task where documenters could benefit of infrastructural support. In addition to the general need of more support in process documentation [37], eliciting AI use oriented documentation requires that documenters focus on producing a best possible account of their own practices instead of trying to adhere to a set of guidelines. While humans generally have difficulties to explain their actions and rationales in detail, it is still much easier than hypothesising what others might need to know about them. Assessing the sufficiency of documentation is also likely to be easier when it is done from a specific, to the documenter familiar, perspective rather than approached as a measure without an explicit point of reference. Furthermore, articulating the perspective makes it explicit for those consulting the documentation, simultaneously helping them to make sense of it.

5. Conclusion

Documenting AI use in humanities research and conveying such documentation is critical for understanding research processes, their direct outputs from research data to diverse forms of publications and broader outcomes. The fast-evolving AI techniques and their uses mean that the adequacy of documentation is a fast-moving target. There is an urgent need to develop a better understanding how to document AI use in humanities research, what needs to be documented and how documentation and the work of documenting should be incorporated in and taken into account in the development of humanities research infrastructures. Major challenges in the process include how to capture enough of the human side of the AI use processes and how to decide what is sufficient documentation. Rather than trying to speculate on future needs, an approach worth considering is to switch perspectives to focus on documenting the transformation AI is used to achieve rather mere interactions with a particular system, and to reflect and document what the documentation itself is attempting to achieve, for whom and how.

Acknowledgements

This work has been supported by the European Research Council (ERC) Grant number 818210 and InterPARES Trust AI project (SSHRC of Canada).

References

[1] I. Grossmann, M. Feinberg, D. C. Parker, N. A. Christakis, P. E. Tetlock, W. A. Cunningham, AI and the transformation of social science research, Science 380 (2023) 1108–1109. doi:10.1126/science.adi1778.

[2] M. E. Kaminski, Regulating the Risks of AI, Boston University Law Review 103 (2022) 1347–1411. doi:10.2139/ssrn.4195066.

[3] S. Cameron, P. C. Franks, I. Huvila, N. Mooradian, Navigating accountability: The role of paradata in AI documentation and governance, Journal of Documentation 81 (2025) 906–926. doi:10.1108/JD-01-2025-0009.

[4] A. Prescott, Bias in Big Data, Machine Learning and AI: What Lessons for the Digital Humanities?, Digital Humanities Quarterly 017 (2023).

[5] A. Foka, L. Eklund, A. S. Løvlie, G. Griffin, Critically assessing AI/ML for cultural heritage: Potentials and challenges, in: Handbook of Critical Studies of Artificial Intelligence, Edward Elgar Publishing, 2023, pp. 815–825.

[6] T. A. Brereton, M. M. Malik, M. Lifson, J. D. Greenwood, K. J. Peterson, S. M. Overgaard, The Role of Artificial Intelligence Model Documentation in Translational Science: Scoping Review, Interactive Journal of Medical Research 12 (2023) e45903. doi:10.2196/45903.

[7] F. Königstorfer, A comprehensive review of techniques for documenting artificial intelligence, Digital Policy, Regulation and Governance 26 (2024) 545–559. doi:10.1108/DPRG-01-2024-0008.

[8] I. Huvila, A Paradata Reference Model, Cambridge University Press, Cambridge, 2025, pp. 180–210.

[9] I. Huvila, Improving the usefulness of research data with better paradata, Open Information Science 6 (2022) 28–48. doi:10.1515/opis-2022-0129.

[10] Y.-H. Liu, I. Huvila, Methods for Generating and Documenting Paradata, Cambridge University Press, Cambridge, 2025, pp. 75–115.

[11] I. Huvila, Future Directions: Making Paradata Matter, Cambridge University Press, Cambridge, 2025, pp. 211–220.

[12] M. Arnold, D. Piorkowski, D. Reimer, J. Richards, J. Tsay, K. R. Varshney, R. Bellamy, M. Hind, S. Houde, S. Mehta, A. Mojsilovic, R. Nair, K. N. Ramamurthy, A. Olteanu, FactSheets: Increasing trust in AI services through supplier's declarations of conformity, IBM Journal of Research and Development (2019). doi:10.1147/JRD.2019.2942288.

[13] E. Frontoni, M. Paolanti, T. P. Lauriault, M. Stiber, L. Duranti, A.-M. Muhammad, Trusted Data Forever: Is AI the Answer?, in: M. Ramanath, T. Palpanas (Eds.), Proceedings of the Workshops of the EDBT/ICDT 2022 Joint Conference, CEUR-WS.org, Edinburgh, 2022, p. paper 1. arXiv:2203.03712.

[14] J. Edmond, J. Lehmann, Digital humanities, knowledge complexity, and the five äporiasöf digital research, Digital Scholarship Humanities 36 (2021) ii95–ii108.

[15] F. Königstorfer, S. Thalmann, Software documentation is not enough! Requirements for the documentation of AI, Digital Policy, Regulation and Governance 23 (2021) 475–488. doi:10.1108/DPRG-03-2021-0047.

[16] F. Königstorfer, S. Thalmann, AI Documentation: A path to accountability, Journal of Responsible Technology 11 (2022) 100043. doi:10.1016/j.jrt.2022.100043.

[17] R. Isdahl, O. E. Gundersen, Out-of-the-Box Reproducibility: A Survey of Machine Learning Platforms, in: 2019 15th International Conference on eScience (eScience), 2019, pp. 86–95. doi:10.1109/eScience.2019.00017.

[18] R. M. Srinivasan, E. Denton, J. J. Famularo, N. Rostamzadeh, F. Diaz, B. Coleman, Art sheets for art datasets, in: Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks, 2021.

[19] H. Semmelrock, T. Ross-Hellauer, S. Kopeinik, D. Theiler, A. Haberl, S. Thalmann, D. Kowald, Reproducibility in machine-learning-based research: Overview, barriers, and drivers, AI Magazine 46 (2025) e70002. doi:10.1002/aaai.70002.

[20] I. Huvila, L. S. Sinnamon, When data sharing is an answer and when (often) it is not: Acknowledging data-driven, non-data, and data-decentered cultures, Journal of the Association for Information Science and Technology 75 (2024) 1515–1530. doi:10.1002/asi.24957.
[21] I. Huvila, L. Andersson, O. Sköld, Patterns in paradata preferences among the makers and reusers of archaeological data, Data and Information Management 8 (2024) 100077. doi:10.1016/j.dim.2024.100077.
[22] I. Huvila, L. Andersson, Z. Friberg, Hs. Liu, O. Sköld, Paradata: Documenting Data Creation, Curation and Use, Cambridge University Press, Cambridge, 2025. doi:10.1017/9781009366564.
[23] O. Azeroual, T. Koltay, Research information in the light of artificial intelligence: Quality and data ecologies, 2024. doi:10.48550/arXiv.2405.12997. arXiv:2405.12997.
[24] T. Gebru, J. Morgenstern, B. Vecchione, J. W. Vaughan, H. Wallach, H. D. III, K. Crawford, Datasheets for datasets, CACM 64 (2021) 86–92. doi:10.1145/3458723.
[25] O. E. Gundersen, Y. Gil, D. W. Aha, On Reproducible AI: Towards Reproducible Research, Open Science, and Digital Scholarship in AI Publications, AI Magazine 39 (2018) 56–68. doi:10.1609/aimag.v39i3.2816.
[26] M. Mitchell, S. Wu, A. Zaldivar, P. Barnes, L. Vasserman, B. Hutchinson, E. Spitzer, I. D. Raji, T. Gebru, Model Cards for Model Reporting, in: Proceedings of the Conference on Fairness, Accountability, and Transparency, ACM, Atlanta GA USA, 2019, pp. 220–229. doi:10.1145/3287560.3287596.
[27] S. Ames, L. Havens, Exploring National Library of Scotland datasets with Jupyter Notebooks, IFLA Journal (2022). doi:10.1177/03400352211065484.
[28] M. Orzechowski, Ł. Opioła, I. L. Martínez, M. Ioannides, P. N. Panayiotou, Ł. Dutka, R. G. Słota, J. Kitowski, Integrated data, metadata, and paradata management system for 3D Digital Cultural Heritage objects: Workflow automation, federated authentication, and publication, Future Generation Computer Systems 174 (2025) 107964. doi:10.1016/j.future.2025.107964.
[29] A. Juneström, I. Huvila, Categorizing methods and approaches for generating and identifying paradata, Journal of Documentation (forthcoming). doi:10.1177/09610006251342811.
[30] I. Huvila, L. Andersson, O. Sköld, Researchers' data processing descriptions—Understanding paradata creation practices and their underpinning instrumentalities, Journal of the Association for Information Science and Technology (2025). doi:10.1002/asi.70003.
[31] L. Börjesson, I. Huvila, O. Sköld, Information needs on research data creation, Information research 27 (2022) isic2208. doi:10.47989/irisic2208.
[32] H. Lönnqvist, The research processes of humanities scholars, in: E. D. Garten, D. E. Williams, J. M. Nyce, S. Talja (Eds.), Advances in Library Administration and Organization, Emerald, Bingley, 2007, pp. 175–202.
[33] M. Bates, D. N. Wilde, S. Siegfried, Research practices of humanities scholars in an online environment: The Getty Online Searching Project report no. 3, Library and Information Science Research 17 (1995) 5–40. doi:10.1016/0740-8188(95)90003-9.
[34] T. Cook, Evidence, memory, identity, and community: Four shifting archival paradigms, Archival Science 13 (2013) 95–120. doi:10.1007/s10502-012-9180-7.
[35] I. Huvila, L. Andersson, O. Sköld, Concluding Discussion: Paradata for Information and Knowledge Management, in: I. Huvila, L. Andersson, O. Sköld (Eds.), Perspectives on Paradata: Research and Practice of Documenting Process Knowledge, Springer International Publishing, Cham, 2024, pp. 249–264. doi:10.1007/978-3-031-53946-6_14.
[36] S. Cameron, P. Franks, B. Hamidzadeh, Positioning Paradata: A Conceptual Frame for AI Processual Documentation in Archives and Recordkeeping Contexts, Journal on Computing and Cultural Heritage 16 (2023) 1–19. doi:10.1145/3594728.
[37] Z. Friberg, I. Huvila, Methods for managing paradata, in: I. Huvila (Ed.), Paradata: Documenting Data Creation, Curation and Use, forthcoming.

Boosting up the Sentiment Analysis Models' Accuracy by Blending Multi-label Learning with a Large Sentiment Lexicon

Dimitrios Kokkinakis[1]

[1] *University of Gothenburg, Box 200, 405 30, Gothenburg, Sweden*

Abstract

This study compares sentiment analysis approaches for Swedish texts using a manually annotated gold-standard dataset. Two methods were examined: i) a multi-label sentiment classifier trained for Swedish, and ii) the Swedish version of VADER, a lexicon-based tool that computes sentiment scores from a vocabulary of polarity-weighted words. The analysis also examined agreement and disagreement between the two methods, with a focus on mixed or context-dependent sentiment. Results indicate that the multi-label classifier aligns more closely with human judgments, especially for medium- or long-text segments with complex or subtle emotional tones. VADER, while prone to errors in idiomatic or nuanced expressions, performs reliably on short, informal utterances, offering computational efficiency and transparency. A hybrid approach combining classifier predictions with lexicon-based scores was investigated to leverage their complementary strengths. Findings underscore the value of rigorous evaluation against human annotations and highlight strategies to improve sentiment analysis in under-resourced languages such as Swedish.

Keywords

sentiment analysis, multi-label classifier, multi-class model, lexicon-based method (VADER)

1. Introduction

Sentiment analysis is a central task in Natural Language Processing (NLP) with broad applications in social media, customer feedback, and automated content evaluation. Despite recent advances driven by machine learning and large language models, lexicon-based methods remain relevant, especially for analyzing short texts or when annotated data is limited. The reason that lexicon-based sentiment methods remain valuable is because they contribute to *transparency* (each word's contribution is explicit, unlike LLMs' opaque decision-making), *adaptability* (they can be easily adapted with custom lexicons for specific domains without retraining or large datasets), and *practicality* (suitable for real-time or resource-constrained settings) — making them complementary to large language models (LLMs) rather than obsolete. This study addresses sentiment analysis for Swedish texts by evaluating two complementary approaches: a multi-label sentiment classifier and a lexicon-based tool (VADER; "Valence Aware Dictionary and sEntiment Reasoner"; [1]). Using a manually annotated gold-standard dataset, we assess their performance in terms of, among other metrics, accuracy, precision, recall, and F1-score, while also leveraging lexicon scores to enhance classifier predictions. Results indicate that the classifier is more effective for longer or syntactically complex sentences, whereas the lexicon-based method better captures mixed or context-dependent sentiments and shorter sentences.

Building on these complementary strengths, we propose a hybrid strategy that integrates both approaches, leading to improved robustness and accuracy for sentiment analysis in under-resourced languages. Previous research in English has explored combining lexicon-based sentiment analysis with machine learning or transformer-based models [2]; such studies demonstrate that hybrid methods can mitigate individual model limitations and improve robustness. The present work builds on this idea, applying and evaluating a similar approach in Swedish, a language with more limited NLP resources.

The rest of the paper is organized as follows: Section 2 describes the dataset and resources; Section 3 details the methodology; Section 4 presents the results and discusses future research directions.

Huminfra Conference 2025, Stockholm, 12-13 November 2025.

dimitrios.kokkinakis@svenska.gu.se (Dimitrios. Kokkinakis)

2. Dataset and linguistic resources

To enable a rigorous evaluation of sentiment analysis methods for Swedish texts, we employed a manually annotated gold-standard dataset as the primary benchmark. This dataset consists of short textual units [*n=2017*], such as social media posts, blogs, passages from Swedish newspapers, and some user-generated comments, each labelled with the standard sentiment categories *positive*, *neutral* and *negative*. The original dataset and annotation process can be found in [3], some minor corrections [n≈20] and adjustments were imposed, after manual inspection, to increase the reliability of the dataset[2]. For instance, sentiment annotation was changed for some data entries: *Hur fina vänner är inte det?* eng. "*What great friends, aren't they?*" from negative in the original gold standard, to positive; *Det blev en pizza och en god vattenmelon-juice istället men det var trevligt.* eng. "*It ended up being a pizza and a tasty watermelon juice instead, but it was nice.*" from neutral in the gold standard, to positive. Moreover, several duplicate entries were removed, ensuring that only unique records remained, e.g. *Kaos på stand up-klubben igår*. eng. "*Chaos at the stand up club yesterday*". The dataset is rather balanced, and Figure 1 (left) shows the distribution of the entries with respect to the three sentiment classes.

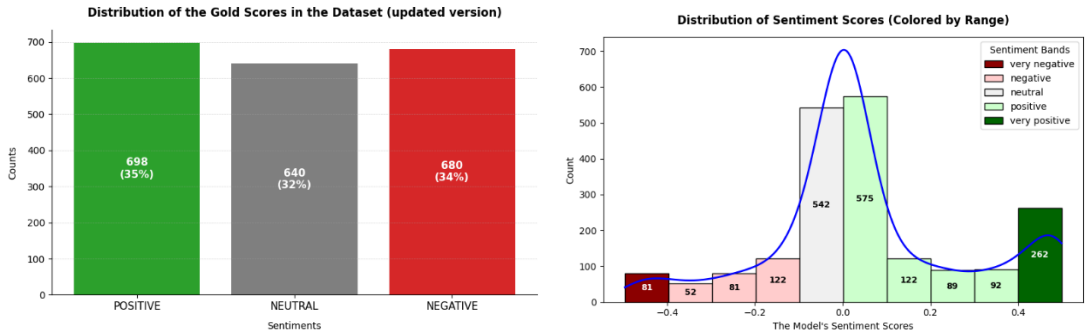


Figure 1: Distribution of the gold dataset's sentiments – the updated version (left) and the distrbution of the KBLab's model output in this same dataset (right) – in this plot, the "very positive" (>= 0.48) and "very negative" (<= -0.48) values are based basically on the obtained scores – *not* highlighted explicitely by the model's output as "very positive" or "very negative".

For the initial annotation we use the '`robust-swedish-sentiment-multiclass`' model from the National Library of Sweden (KBLab[3]). According to the creators of the model [4], the model is a release of a robust, multi-label sentiment classifier finetuned on Megatron-BERT-large-165K. The model was trained on approximately 75K Swedish texts from multiple linguistic domains and datasets. The model addresses gaps in Swedish sentiment analysis by including a neutral category and training across diverse datasets beyond reviews. Using data from reviews, Twitter, news, immigration discourse, and translated texts, it achieves strong generalization and accuracy (0.80 multiclass, 0.88 binary), making it more robust than earlier Swedish sentiment models. In addition, we applied a lexicon-based approach on the output of the previous model. Specifically, the VADER sentiment lexicon, adapted and extended for Swedish (svVADER; [5]), containing polarity scores associated with words and multiword expressions, was used. VADER ignores word context, especially when word order or distant lexical items intervene in multi-word expressions. Nonetheless, VADER has a negation identification

[2] The column *sentimentannotation.csv* from the original gold standard file: https://raw.githubusercontent.com/richi1998/Evaluating-Lexicon-Based-Models-versus-BERT-for-Sentence-Level-Sentiment-Ana-lysis-in-Swedish/refs/heads/main/koden/**Data_svm.csv** was used for the evaluation exercise. It was manually reviewed, slightly updated, and subsequently used in the experiment. The resulting gold file, renamed *updatedGoldDataset.csv* can be found here: https://github.com/DimitrisKokkinakis/swedish-notebooks/blob/main/textual-resources/HiC-2025/**updatedGoldDataset.csv**.

[3] KBLab's blog post (https://kb-labb.github.io/posts/2023-06-16-a-robust-multi-label-sentiment-classifier-for-swedish/) describe the model as "multi-label" because the underlying architecture could assign multiple sentiment labels simultaneously. However, the training data and released checkpoints use single-label annotations, and the model outputs one dominant class per text (i.e., the highest-probability label). Therefore, the classifier is a multi-class model, predicting one sentiment label per sentence (positive, negative, or neutral), rather than a true multi-label setup.

mechanism to shift polarity under certain circumstances. The enhanced Swedish lexicon[4] includes more than 50,000 entries, encompassing several thousand multi-word expressions, compared to 5,501 entries in the original translation[5] provided in [7].

Together, these resources provide the foundation for both the learning approach (multi-label classifier) and the knowledge-based strategy (lexicon-driven sentiment scoring). They also allow us to explore hybrid methods that combine lexical information with model-based predictions, thereby addressing the limitations of working with under-resourced languages.

3. Methodology, experimental design and results

Performance was quantified using various standard metrics for sentiment analysis evaluation. These metrics collectively provide a comprehensive assessment of classification performance. Accuracy offers an overall measure of correctness, while the Matthews Correlation Coefficient (MCC) captures the balance between true and false classifications, making it particularly informative under class imbalance; however, this metric is less informative in the present case, as the dataset is relatively balanced, but it was included for completeness and comparability with related studies. Precision and Recall quantify, respectively, the proportion of correctly identified positive predictions and the ability to retrieve all relevant instances, with their harmonic mean expressed as the F1 Score. The macro-averaged metrics treat each class equally, reflecting performance across categories regardless of frequency, whereas the micro-averaged metrics aggregate all instances to emphasize overall system performance relative to the gold-standard annotations (see Figure 3).

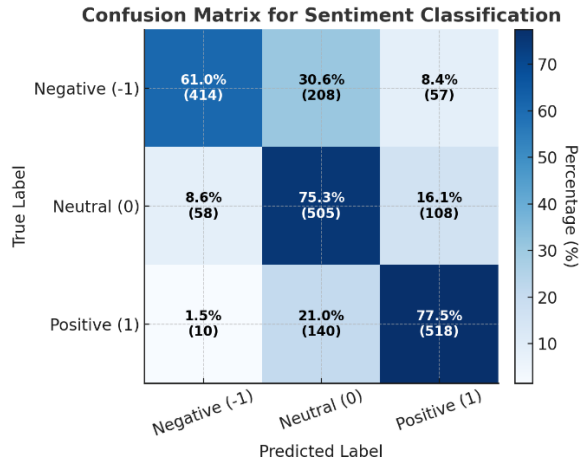


Figure 2: The confusion matrix of the KBLab's sentiment classifier

Figure 2 on the left shows clearly how the KBLab sentiment classifier performed across the three sentiment categories. The darker diagonal cells indicate stronger agreement with the gold standard, while the lighter off-diagonal cells highlight areas of confusion (especially between neutral and neighboring classes). The system achieves roughly 72% accuracy with a macro-F_1, indicating a balanced though moderate performance across the three sentiment classes. Precision is highest for negative (0.86) and positive (0.76) sentiments, while neutral (0.59) lags behind, suggesting that the model tends to confuse neutral expressions with polar ones—a typical challenge in multi-class sentiment analysis.

The two resources were applied sequentially, first the KBLab model followed by the svVADER on the dataset entries (rows) where the model did not agree with the gold standard. On the specific gold standard dataset, the model's evaluation metrics were *accuracy* 72% and *MCC* 59.04%. svVADER was applied on the 565 mismatches (rows), i.e. the cases in which the model and gold standard did not agree, and the results on this subset's evaluation metrics were *accuracy* 52.38% and *MCC* 29.69%. From the number of mismatches, 296 rows were assigned the correct sentiment label and, while 269 were assigned an erroneous sentiment. The combined, global accuracy was 86.67% and the MCC 80.18%.

[4] Selected subsets of the svVADER's lexicon have been evaluated using LLMs (ChatGPT) with manual follow-up. Consistent with similar studies [6], ChatGPT performed quite well, suggesting LLMs can effectively support initial annotations and accelerate lexicon development. One of such subset evaluations focused on entries containing the substring 'under', such as ***under****bart* eng. wonderful; *väl****under****byggd* eng. well-founded and ***under****känd* eng. failed (n=250) can be found here: https://github.com/DimitrisKokkinakis/swedish-notebooks/blob/main/textual-resources/HiC-2025/svVADER-vs-LLM-proofOfConcept.xlsx.

[5] https://github.com/marcusgsta/vaderSentiment/tree/master/vaderSentiment (visited 2025-10-27).

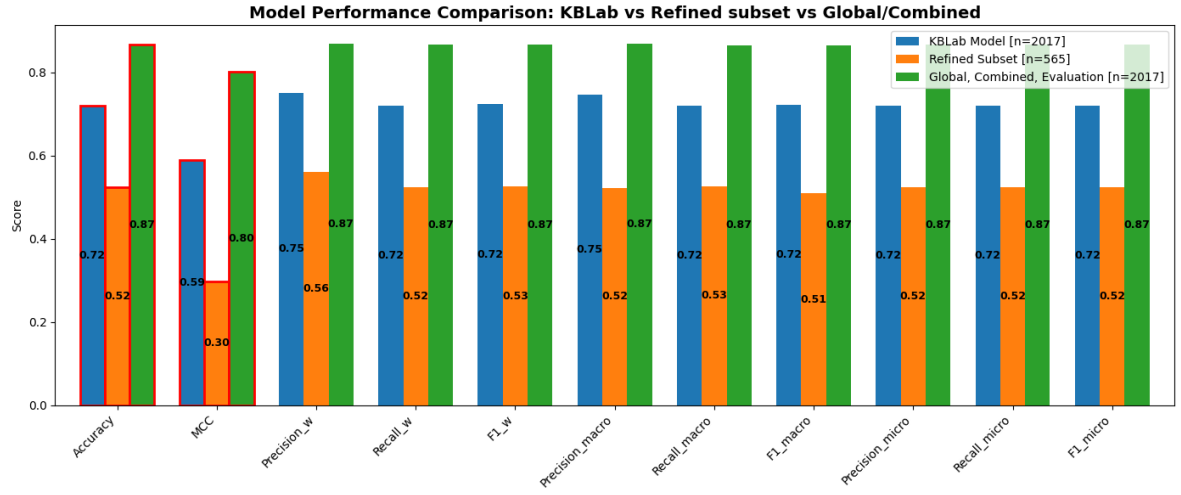


Figure 3: The various metrics used for the evaluation: the KBLab model (the left bar in each bar group), the refined subset, that is the datapoints erroneously annotated by the KBLab model (the middle bar in each bar group) and the combination scores of the two (the right bar in each bar group).

The final mismatched cases (n = 269) predominantly originated from sentences containing long, coordinated constructions anchored by a negative lexical element, which often resulted in scope-related interpretation errors. A further source of misclassification involved brief sentences whose correct sentiment interpretation depended on contextual cues or world knowledge beyond the textual input. Additional inconsistencies were observed in instances featuring figurative, metaphoric, or ironic expressions, where the intended evaluative meaning was challenging for the model to discern. Representative examples of these categories are provided in Table 1.

Table 1
Examples of Sentiment Annotation Challenges

Category	Swedish sentence	English glossing	Description/ Challenge	Gold Annotations
Long coordinated constructions (scope errors)	Bristen på värme, medmänsklighet, trygghet och tröst är total.	The lack of warmth, humanity, security, and comfort is total.	Sentence anchored by a negative word ('bristen på'). Complex coordination leads to potential scope interpretation errors.	–1 (Negative)
Very short input sentences requiring world knowledge & contextual disambiguation	År 1958 kom Sverige tvåa. *or* Det blev ett brons för henne i big air med skidor.	In 1958, Sweden came second. *or* She won a bronze medal in big air with skis.	Short sentences needing external or contextual understanding (e.g., sports results).	+1 (Positive)
Metaphoric sentences	Se hur lång tid det tar innan maskinen äter ditt kort.	See how long it takes before the machine eats your card.	Figurative use of 'eats' (metaphor); potential for literal misinterpretation.	–1 (Negative)
Sarcastic or ironic sentences	Det var någon slags alkoholist-bingo typ.	It was some kind of alcoholic bingo type.	Sarcastic or ironic tone is usually difficult for models to detect.	–1 (Negative)

The motivation for applying the lexicon-based model *only* to sentences where the first model fails is to explore their complementary strengths. This approach helps reveal where each method performs better—for instance, the lexicon model may handle explicit polarity words more effectively, while the machine learning model captures contextual nuances. Although such an approach is not directly applicable in real-world settings where true labels are unknown, it serves as a *proof of concept* demonstrating the potential of selective combination. In practice, this insight could inform confidence-

based or ensemble strategies, where the lexicon model acts as a fallback when the main model's confidence is low.

4. Conclusions and future work

The paper evaluates Swedish sentiment analysis tools — svVADER and the KBLab sentiment model. The results are valuable given the limited Swedish NLP resources. It also proposes a hybrid approach combining both methods to offset their individual weaknesses, a promising idea for other low-resource languages. The hybrid accuracy scores were calculated by sequentially combining the outputs of two sentiment analysis systems—a transformer-based model and a lexicon-based method—and then evaluating their cumulative performance relative to a gold-standard dataset. In essence, the hybrid accuracy scores represent a cumulative metric reflecting the complementary strengths of both models—the contextual robustness of the transformer-based classifier and the lexical sensitivity of the rule-based system—applied in a corrective, sequential manner.

This study has shown that sentiment analysis for Swedish texts can be substantially improved by combining multi-label classification with lexicon-based methods. The multi-label classifier aligned more closely with human annotations, particularly for complex or ambivalent sentences, while the lexicon-based approach contributed transparency and efficiency, capturing nuances in short ones. By integrating both approaches, we achieved a hybrid system with notably higher accuracy (86.67%) than either method independently. These findings confirm that leveraging complementary strengths is especially valuable in under-resourced language contexts, where annotated data remains limited.

Despite the promising results, several challenges remain. The analysis revealed recurrent difficulties in handling coordinated constructions, context-dependent expressions, sarcasm, and irony—phenomena that continue to challenge both statistical and lexicon-driven approaches. Furthermore, reliance on static lexical resources makes it difficult to adapt to emerging vocabulary and evolving usage in social media and digital communication. Future work will address these limitations in several directions. First, expanding the gold-standard dataset with broader domain coverage and richer annotations will improve both training and evaluation. Second, integrating contextual embeddings from large-scale transformer models could enhance the detection of subtle sentiment cues, such as sarcasm [8; 9], irony [10] or metaphors [11]. Third, adaptive or dynamically updated lexicons may mitigate the rigidity of current dictionary-based resources. Finally, applying the hybrid strategy to other under-resourced languages will test its generalizability and contribute to cross-linguistic sentiment analysis research. In sum, this work provides both methodological insights and practical contributions toward more robust, accurate, and interpretable sentiment analysis systems for Swedish and beyond.

Acknowledgements

Work on the article has been supported by The National Language Bank of Sweden (Nationella Språkbanken) and HUMINFRA, the Swedish national infrastructure supporting digital and experimental research in the Humanities and their participating partner institutions, both funded by the Swedish Research Council (2018–2024, contract 2017-00626; 2022–2024, contract 2021-00176).

References

[1] C. Hutto, E. Gilbert. Vader: a parsimonious rule-based model for sentiment analysis of social media text. In: Proceedings of the international AAAI Conference on Web and Social Media, vol. 8, 2014, pp. 216–225.

[2] L. Barros, A. Trifan, and J. L. Oliveira. VADER meets BERT: sentiment analysis for early detection of signs of self-harm through social mining. In: CLEF 2021 – Conference and Labs of the Evaluation Forum, Bucharest, Romania, 2021. https://ceur-ws.org/Vol-2936/.

[3] R. Mansour, E. Nilsson. Evaluating Lexicon-Based Models versus Bert for Sentence Level Sentiment Analysis in Swedish, 2024, https://github.com/richi1998/Evaluating-Lexicon-Based-Models-versus-BERT-for-Sentence-Level-Sentiment-Ana-lysis-in-Swedish

[4] H. Hägglöf, A Robust, Multi-Label Sentiment Classifier for Swedish. June 16, 2023. https://huggingface.co/KBLab/robust-swedish-sentiment-multiclass
[5] D. Kokkinakis, R. Muñoz Sánchez, and Mia-Marie Hammarlin. Scaling-up the Resources for a Freely Available Swedish VADER (svVADER) in: Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa), Tórshavn, Faroe Islands. University of Tartu Library, 2023, pp. 667–672.
[6] F.S. Marcondes, A. Gala, M. Rodrigues, J.J. Almeida, P. Novais. Lexicon Annotation with LLM: A Proof of Concept with ChatGPT. In: Quintián, H., et al. Hybrid Artificial Intelligent Systems. HAIS 2024. Lecture Notes in Computer Science(), vol 14858. Springer, Cham. https://doi.org/10.1007/978-3-031-74186-9_16
[7] M. Gustafsson. Sentiment analysis for tweets in Swedish: Using a sentiment lexicon with syntactic rules. Bachelor's thesis. [Online]. Linnaeus University, Sweden. 2020. https://www.diva-portal.org/smash/record.jsf?pid=diva2%3A1391359&dswid=-8277
[8] E. Riloff, A. Qadir, P. Surve, L. de Silva, N. Gilbert, and R. Huang. Sarcasm as Contrast between a Positive Sentiment and Negative Situation in: Proceedings of the Empirical Methods in Natural Language Processing (EMNLP), Seattle, Washington, USA Association for Computational Linguistics, 2013, pp. 704–714. https://aclanthology.org/D13-1066/
[9] Q. Li, Z. Li, W. Liu, X. He, and Y. Pan. Sarcasm-GPT: advancing sarcasm detection with large language models, The Computer Journal, 2025; bxaf055, https://doi.org/10.1093/comjnl/bxaf055
[10] B. Ghanem, J. Karoui, F. Benamara, P. Rosso, and V. Moriceau. Irony Detection in a Multilingual Context. Advances in Information Retrieval. 2020 Mar 24;12036:141–9. https://dl.acm.org/doi/10.1007/978-3-030-45442-5_18.
[11] S. Yang, D. Zhang, J. Ren, Z. Xu, X. Zhang, Y. Song, H. Lin, and F. Xia. Cultural Bias Matters: A Cross-Cultural Benchmark Dataset and Sentiment-Enriched Model for Understanding Multimodal Metaphors, in: Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (ACL), Vienna, Austria. Association for Computational Linguistics. 2025, pp. 26301–26317. https://aclanthology.org/2025.acl-long.1275/.

Mapping Soundscapes of Warning

Experimental Interfaces for Public Sound Culture

Johan Malmstedt[1,*], Kirill Mitsurov[2] and Marie Cronqvist[3]

[1]*Linköping University (LiU), Sweden*
[2]*C²DH (Centre for Contemporary and Digital History), University of Luxembourg*
[3]*Linköping University, Sweden*

Abstract

This short paper introduces *Soundscapes of Warning*, an experimental research application designed to support the comparative study of public warning signals as cultural and aesthetic artefacts. Developed through a collaboration between Linköping University and C²DH at the University of Luxembourg, the platform enables users to explore how alarm sounds, sirens and civil alert signals, vary across national and historical contexts. By combining geographic comparison with custom-designed 3D visualizations of alarm signals, the application offers a new model for investigating how warnings and urgency have been rendered sonically in different societies. Instead of approaching warning sounds as purely functional or technical signals, the platform emphasizes their role in shaping public space and the semiotics of danger. Designed as both a research tool and an interpretive interface, *Soundscapes of Warning* contributes to current efforts in the digital humanities to critically engage with sound as a mediated and historically contingent form.

Keywords

digital humanities, research infrastructure, warning signals, sound studies

1. Introduction

At the instant a warning signal echoes across its surroundings, it carries something both unsettling and familiar. Designed to announce disruption, the alarm also resonates like a natural phenomenon: ubiquitous and habitual. Over time, these sounds become neutralized within the modern soundscape, blurring the line between infrastructure and atmosphere.

But what does it mean to hear danger? And how does the sonic experience of warning differ between societies? Alarms are never merely technical signals; they are culturally encoded events that shape, and are shaped by, aural culture and the history of science. For some, sirens signal protection; for others they evoke trauma or failed systems. Scholars in sound studies have emphasized how public sound not only produces listeners, but also distributes agency and meaning across the soundscapes of modernity [1, 2].

This paper introduces the *Soundscapes of Warning* application, an experimental tool that seeks to defamiliarize alarms by treating them as cultural and aesthetic artefacts. Rather than natural sounds, they are understood as products with specific histories that reveal how societies construct the semiotics of warning. The project presented in this paper is a part of a larger project titled ”Soundscapes of Warning: The Past, Present and Futures of Viktigt Meddelande till Allmänheten (VMA)” (Vetenskapsrådet VR-SÄKER, 2023-05736, PI Marie Cronqvist) and which interrogates the cultural and infrastructural history of the Swedish air-raid siren.

2. The Cultural History of Sound Alarms

Before their deployment as instruments of civil defense, sirens were woven into the industrial soundscape: they punctuated the rhythms of factory labor, guided ships through fog, and signaled the working day

Huminfra Conference 2025, Stockholm, Sweden.
*Corresponding author.
johan.malmstedt@gu.se (J. Malmstedt)
https://www.uni.lu/c2dh-en/people/kirill-mitsurov/ (K. Mitsurov)

across towns and cities [3]. During the Second World War and the Cold War, their howls became synonymous with looming aerial attacks and the vulnerability of civilian life. Today, however, many of these mechanical warning systems are being supplanted by digital and electronic solutions, rendering the wail of the siren increasingly a historical phenomenon—an acoustic trace of modernity that calls for cultural and historical interpretation.

The *Soundscapes of Warning* project continues and enriches this inquiry into sirens as cultural sounds. Drawing on Marie Cronqvist's work on the acoustemology of sirens in Sweden [4], it treats alarms as ways of knowing and governing through sound: signals that not only announce danger but also shape listeners, encode authority, and structure civic routines. We suggest to consider these sounds, not as neutral tones, but as historically and socially embedded artefacts. The project develops tools to make these sonic infrastructures comparable and explorable, combining historical research with experimental methods to understand how societies organize attention, memory, and vulnerability through sound.

3. Siren Scholarship

In *Sirens*, Michael Bull observes that the exclusion of modern alarms from critical discourse is striking; even within sound studies, warning sirens are often treated only in passing despite their central role in twentieth-century public life [5]. Much scholarship has instead focused on military sound, or what Steve Goodman in *Sonic Warfare* (2009) describes as the weaponization of vibration and noise—sound as attack [6]. By comparison, alarms and sirens—sound as defence—have often been neglected, even though their regular tests and sudden activations leave deep imprints on public memory.

More recently, artistic and scholarly projects have begun to foreground sirens as cultural sounds in their own right. Inquires into the early history of German alarm system have been reoccurring topic ([7], [8]). Whereas Aura Satz's *Preemptive Listening* (2018) and its associated roundtables frame them as instruments of sonic governance—technologies that summon publics, produce listening subjects, and delineate lines of protection and exposure [9].

The *Soundscapes of Warning* application builds on these insights by treating sirens as historically and culturally embedded artefacts. Drawing on acoustemological perspectives [4], it examines how alarms encode authority, vulnerability, and civic routines. By combining historical research with experimental visualization, it explores sirens not as neutral signals but as infrastructures of governance and collective attention.

4. Design and Development

The application is designed to estrange and reframe the auditory character of warning signals. Rather than treating alarms as functional audio, it presents them as historically shaped media forms with aesthetic, political, and cultural dimensions. The aim is to denaturalize warning sounds and allow users to encounter them as artefacts—both data and cultural record.

Each signal is rendered as a volumetric 3D structure using audio feature extraction. Building on earlier experimental audio analysis methods [10], these objects are interpretive rather than reconstructive: tangible, printable, and interactive forms that reframe short signals in ways once applied to extended broadcasts.

The development of the *Soundscapes of Warning* application also connects to ongoing interface experiments at the C^2DH. In particular, it builds on design strategies explored in the *3D Stories* project, which combined interactive models with narrative annotation to make historical artefacts accessible in new ways. Like that project, our platform uses three-dimensional forms not only for display but as vehicles for interpretation, linking cultural artefacts to broader research narratives. In this sense, *Soundscapes of Warning* represents a branch of this design lineage, extending narrative-3D methods from material history into the domain of public sound culture.

Built with a modern web architecture, the *Soundscapes of Warning* application is implemented in

TypeScript and React, bundled with the Vite framework for fast rendering and modular development. The interface runs entirely in the browser using pre-rendered audio data, ensuring stable performance and reproducibility. Visualizations are generated from structured datasets that can, in future iterations, connect to real-time audio APIs if formatted accordingly. The current build is hosted on the C^2DH servers (uni-c2dh.lu), following initial testing through Netlify, and the open-source repository allows others to fork and adapt the code for their own research interfaces. Designed as a framework rather than a fixed platform, it enables the generation of three-dimensional soundscapes from processed audio, parameter adjustment through frequency and amplitude filters, and export of models for analysis or physical fabrication.

4.1. Interface and Interaction Design

Figure 1: Opening globe interface of the *Soundscapes of Warning* application.

On entry, users encounter a globe that anchors sounds in geographic context. By zooming and clicking pinned locations, they access recordings and metadata, situating alarms as local expressions of sound culture. The interface is rendered directly in JavaScript, allowing for smooth interaction across devices without external software.

The globe reflects the project's broader design identity, which employs minimalist cartographic layouts, monochrome palettes, and restrained motion to question sonic standardization and foreground spatial difference. This visual language is shared across the larger *Soundscapes of Warning* project—appearing in publications, workshops, and installations—ensuring aesthetic continuity between research outputs. In this way, the application serves not only as a technical tool but as part of a wider design framework aimed at defamiliarizing warning signals and encouraging comparative listening.

4.2. Shaping Sound: 3D Visualization of Spectral Data

The core visualization method adapts spectrography into a three-dimensional form: frequency mapped to the Y-axis, time to X, and amplitude to Z. This produces sculptural terrains that make sonic profiles tangible. Developed in the mid-20th century for phonetics [11] and later used in bioacoustics [12] and seismology, spectrography translates the invisible into interpretable form. Following Sterne [2], we also

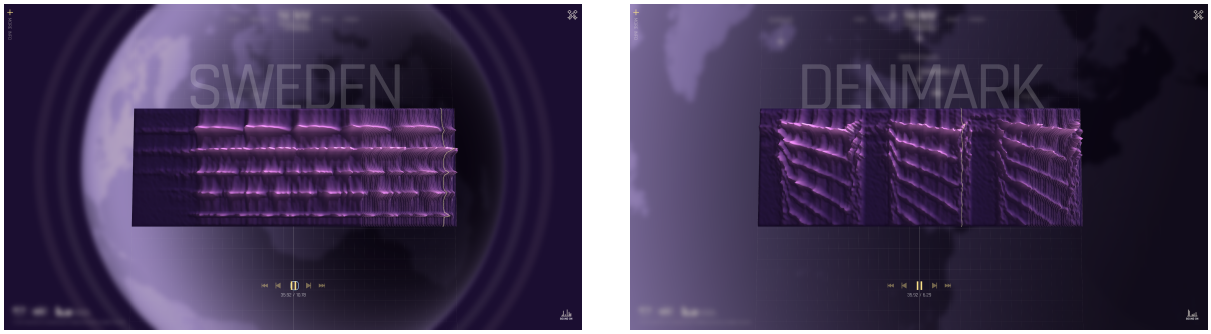


Figure 2: Visualisation of the Swedish and Danish alarm signals displayed next to each other.

recognize that such visualizations shape not only understanding but governance. In our application, spectrograms are rendered interactively in the browser; users can rotate, zoom, and pan the objects, extending listening into looking and interpretation. The shift from spectrogram-as-image to sound-as-shape is both aesthetic and conceptual, prompting reflection on how different signal cultures generate different geometries of alarm.

4.3. Metadata and Comparative Perspectives

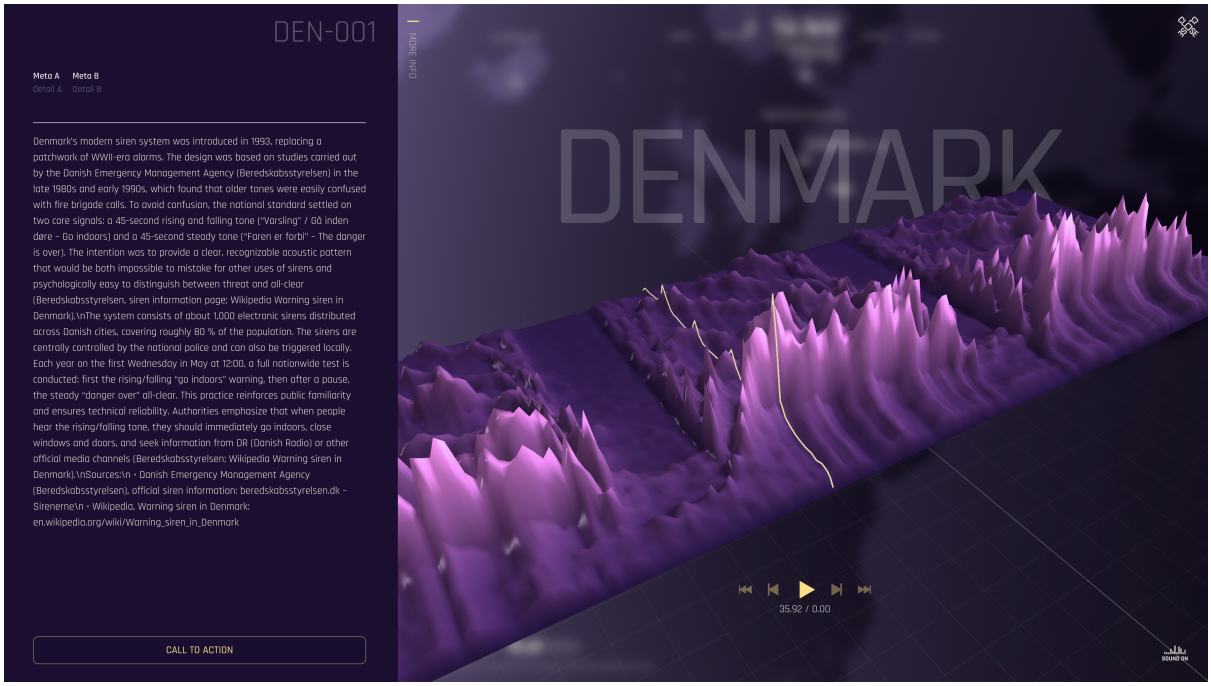


Figure 3: Metadata interface in the *Soundscapes of Warning* application.

Each alarm signal in the *Soundscapes of Warning* app is paired with metadata that situates it within a broader cultural and technical history. For now, this includes contextual notes gathered from national warning webpages, highlighting how systems differ in their codification, testing routines, and intended behaviors. Despite their ubiquity, the comparative history of national warning systems remains underexplored, making this layer of metadata a starting point for future research.

Planned updates will extend metadata to include precise recording locations and details about the devices themselves—such as air horns, rotating sirens, or electronic loudspeakers—since these material differences strongly shape the character of alarm sounds [3, 13]. By linking auditory form with infrastructural context, the platform not only catalogs signals but also opens pathways for comparative study

across nations and technologies.

5. Conclusion

At a time when warning signals are actualized with increasing urgency, it may also be time for scholarship to move siren sounds from the cultural periphery to centre stage. We hope that the *Soundscapes of Warning* application serves this dual role: guiding future research into the local histories of alarm while also opening these sonic infrastructures to a wider public as cultural artefacts. Though still ongoing, we aim to demonstrate how digital humanities approaches can make audible sonic cultures. By enabling comparative exploration, it invites both scholars and lay audiences to reflect on what it means to hear danger and how the semiotics of warning vary across cultural and historical contexts.

References

[1] R. M. Schafer, The Soundscape: Our Sonic Environment and the Tuning of the World, Destiny Books, 1977.

[2] J. Sterne (Ed.), The Sound Studies Reader, Routledge, 2012.

[3] K. Bijsterveld, Mechanical Sound: Technology, Culture, and Public Problems of Noise in the Twentieth Century, MIT Press, 2008.

[4] M. Cronqvist, The acoustemology of sirens: A century of urban communication infrastructures of fear and public sonic warnings in sweden, Paper presented at "20 Years of Sound Environments" conference, Lund University, 2024. Project overview available via the Soundscapes of Warning blog :contentReference[oaicite:1]index=1.

[5] M. Bull, Sirens, The Study of Sound, Bloomsbury Academic, London, 2020. First published February 6, 2020.

[6] S. Goodman, Sonic Warfare: Sound, Affect, and the Ecology of Fear, MIT Press, 2009.

[7] M. Schmidt, Alarm im Äther: Luftschutz in Deutschland 1929–1945, Wallstein Verlag, Göttingen, 2012. A detailed study of air-raid protection and warning infrastructures in Germany.

[8] C. Birdsall, Nazi Soundscapes: Sound, Technology and Urban Space in Germany, 1933–1945, Amsterdam University Press, Amsterdam, 2012. Part of the Sound Studies series.

[9] A. Satz, Preemptive listening, Film (with associated roundtable discussions), 2018.

[10] J. Malmstedt, Sound out of Time: Signal Archaeology of Swedish Public Service Radio, 1980–1999, Ph.D. thesis, Umeå University, Umeå, Sweden, 2024. PhD dissertation in Digital Humanities.

[11] M. Joos, Acoustic Phonetics, Linguistic Society of America, 1948.

[12] D. E. Kroodsma, Song patterns of the eastern phoebe (*Sayornis phoebe*), The Auk 106 (1989) 23–35.

[13] E. Thompson, The Soundscape of Modernity: Architectural Acoustics and the Culture of Listening in America, 1900–1933, MIT Press, Cambridge, MA, 2002.

NER som ett Källidentifieringsverktyg. Erfarenheter av Svenska BERT för Digital Historia 1.25

Jens Norrby[1,2]

[1] *Institutionen för litteratur, idéhistoria och religion, Göteborgs universitet, Renströmsgatan 6, Göteborg, 40530, Sverige*
2 Centre for European Research at the University of Gothenburg (CERGU), Göteborgs universitet, Renströmsgatan 6, Göteborg, 40530, Sverige.

Abstract
The paper explores the experiences of working with Named Entity Recognition (NER) in Swedish parliamentary records. Thus, it provides a practical account of a methodology that employs Swedish BERT and its NER functionality on a historical dataset. It also discusses the relevance of this case to the broader relationship between digital and traditional intellectual history. The study described used NER to identify the geographical areas and placenames within Swedish parliamentary discourse from 1887 to 1914. Taken together, this list of locations could be used to determine the aggregate frequencies of geographical groupings, in this case predominantly nations. The quantitative findings were subsequently used to navigate the data set and identify the most relevant texts for qualitative, contextual close readings. This paper argues that there are strengths in incorporating digital tools within traditional intellectual history in accordance with the principle of 'digital history 1.25'.

Keywords
Named Entity Recognition, Parliaments, Mental Maps, BERT, Digital History

1. Inledning

Vid sidan om den spännande utvecklingen inom nya metoder baserade på stora språkmodeller så kommer jag i detta paper att reflektera kring mitt arbete med ett enklare, mer begränsat, AI-verktyg som ett stöd till traditionell språkbrukshistoria. Inom ramen för projektet "Geography of Turn-of-Century Politics" (Åke Wibergs stiftelse) har jag använt mig av Named Entity Recognition (NER) som en avgörande hjälp i att studera den geografi som förekommer i svensk riksdagsdebatt och de mentala kartor som figurerade i politiska resonemang och argument. Projektet är litet i sin omfattning men har utmynnat i ett artikelutkast om Tysklands centrala roll i riksdagsdebattens internationalisering (*Scandinavian Journal of History*, under granskningsförfarande). Tillsammans med finansiering från Helge Ax:son Jonssons stiftelse och Wahlgrenska stiftelsen har projektet också möjliggjort vidare analys av tidningsmaterial, vilket sammantaget utgör en utgångspunkt för utforskandet av mentala kartor och geografiska platser inom svensk politisk diskurs i en bredare mening.

I detta paper kommer jag att återge mina erfarenheter kring hur den första artikelns resultat möjliggjordes av NER, hur modellen fungerade att använda och vad detta antyder om potentiella användningsområden framåt. Utifrån erfarenheten av att jobba med ett projekt där digitala metoder spelar en avgörande men mycket begränsad roll kommer jag också att kort reflektera kring hur den här studien positionerade sig mellan kvantitativ och kvalitativ forskning.

1.1. Disposition

Kommande avsnitt kommer att ge en översiktsbild av forskningsfältet och en motivering till vilken lucka och behov som studiet av diskursiv geografi kan fylla. Avsnitt 3 ger en beskrivning av hur studien genomfördes praktiskt och vilka metodologiska överväganden som ledde till dess utformning. Avsnitt 4 rör sig vidare från de konkreta, fallspecifika, redogörelserna och lyfter blicken med reflektioner utifrån lärdomar som är relevanta för framtida forskning.

Huminfra Conference 2025, Stockholm, 12-13 November 2025.

jens.norrby@lir.gu.se (J. Norrby)

2. Mentala kartor & diskursiv geografi

Mentala kartor (*mental maps*) har kommit att bli ett väl etablerat begrepp inom studiet av historiska geografiska föreställningar. Teoretiskt grundar sig begreppet i antagandet att en geografisk representation aldrig endast är en återgivning av den fysiska verkligheten, utan en symbolisk representation som framhäver vissa kvalitéer och undangömmer andra – en insikt som brukar åberopas någon av klassikerna *Orientalism* och *Imagined Communities* [1, 2]. Med utgångspunkten att geografisk representation återger både fysiska förhållanden och mentala/symboliska tolkningsramar för dessa så kan begreppet "mentala kartor" benämna två fenomen. Mentala kartor kan för det första åsyfta de mentala aspekterna som kan utläsas från fysiska kartor – alltså de av kartans uttryck som inte endast återger den fysiska verkligheten – och för det andra de mentala representationer av geografiska förhållanden som varje individ utgår från i sin förståelse av världen. Mitt projekt intresserar sig för denna senare kategori, vad Norbert Götz och Janne Holmén benämner som "latenta" mentala kartor och Lauren Benton som geografi "off the map" [3, 4]. Dessa är av idéhistoriskt intresse för att kunna utläsa "kulturellt och historiskt specifika föreställningar om vår erfarbara och föreställningsbara omvärlds rumsliga struktur" (*kulturell und historisch spezifische Vorstellungen von der räumlichen Strukturierung ihrer erfahrbaren und ihrer vorstellbaren Umwelt*) [5, s. 495].

Potentialen hos digitala metoder att utforska spatiala aspekter av historien blev tydligt tidigt i den digitala revolutionen och Geographic Information Systems (GIS) används innovativt för att utforska frågor som berör geografiska föreställningar och mentala kartor [6–8]. Begreppet mentala kartor har också fått fäste inom kvalitativ historisk forskning och studier om historiska geografiska föreställningar [9]. Dessa studier centreras ofta kring enskilda aktörer, för att bättre förstå deras förståelse av komplexa historiska fenomen såsom kolonialism eller geopolitiska hot [10, 11], men begreppet har också tillämpats på de historiska föreställningar som knutits från olika håll till en viss region [12]. Slutligen finns det parallella spår, såsom geografisk begreppshistoria, med intresse för vilka associationer och antaganden som knyts till det geografiska språkets användning, snarare än hur dessa kan byggas samman till en symboliskt meningsfull mental kartbild [13].

Studiet av mentala kartor har generellt intresserat sig för föreställningar om geografiska förhållanden men inte lika mycket för vilka föreställningar som framkommer från geografiska beskrivningar. Jo Guldis artikel "The Official Mind's View of Empire, in Miniature" sticker här ut som en tentativ undersökning av vad förekomsten av olika platsnamn inom en diskurs återspeglar av diskursens geografiska föreställningar [14]. Mitt projekt följer i detta spår och studerar geografin inom svensk riksdagsdebatt som summan av de platsnamns som förekommer, alltså vilken geografi som hänvisades till under en viss period och vilka mönster och förändringar som kan synas i diskursens förhållande till olika geografiska områden.

3. Studiens genomförande

En trolig anledning till att det varit ovanligt att studera sammanställningar av de geografiska hänvisningarna är att det är mycket tidskrävande att manuellt identifiera de platser som förekommer inom en substantiell korpus, såsom svensk riksdagsdebatt. Sedan 80-talet har det dock funnits en disciplin dedikerad till *Named Entity Recognition and Classification* (NERC), där man med digital processorkraft utvecklat metoder för att identifiera namngivna entiteter och tidsangivelser i text, däribland geografi. Principiellt har den geografiska avläsningen (geoparsing) skett antingen genom 1) avstämning gentemot en ortnamnsförteckning, 2) regel-baserad process eller 3) maskininlärning [15]. Den sistnämnda har den senare tiden väckt allt större intresse och det tog inte lång tid efter deras ikoniska artikel 'Attention is All you Need' (2017) för Google att lansera sin Bidirectional Encoder Representations from Transformers (BERT) [16]. BERT är byggd på öppen källkod och visade 2018 upp imponerande prestationer inom en rad olika uppgifter, till exempel NER. 2020 släppte Kungliga Bibliotekets digitala labb (KBLabb) en version som de tränat på svenskt material, vilket öppnade möjligheten för att med hjälp av en stor språkmodell (LLM) identifiera namngivna platser i svenska textkorpus. För detta projekt har modellen bert-base-swedish-cased-ner använts, vilken tränats speciellt för NER på korpuset SUC 3.0 [17].

Jag utgick från det digitaliserade riksdagskorpuset som finansierats inom ramen för projekten Swerik och Westac och utifrån det kunde jag lista alla de ord som modellen kategoriserar som platser mellan 1887 och 1914 [18]. Resultatet blev en lång lista (553 330 hänvisningar, 11 795 unika termer) med potentiellt relevanta ort- och platsnamn. Inom NERC har man funnit olika lösningar för att kombinera denna kvantitativa geografiska avläsning (*geoparsing*) med en kvantitativ geografisk kodning (*geocoding*) där språkliga, potentiellt geografiskt relevanta, uttryck inordnas i ett system kopplat till faktisk geografi. Genom ortnamnsförteckningar, algoritmiska regelverk eller boolesk-statistiska operationer ges varje relevant språkligt uttryck en geografisk korrespondent som bortser från lokala namnvariationer, grammatiska konstruktioner (t.ex. genitivform), historiska förändringar och felstavningar/avläsningsfel. Resultatet är en lista med identifierbara platser som i nästa steg lagras i hanterbart format (*spatial storage*), ibland kureras ytterligare genom rumslig slutledning (*spatial inference*) för att säkerställa inbördes logik, och slutgiltigen visualiseras med GIS-programvara (*application visualization*) [15].

Den här studien involverade inte på någon geografisk visualisering och den geografiska kodningen utfördes kvalitativt genom att slå ihop språkliga variationer (Torneträsk/ Torne träsk/Torneåträsk), OCR-variationer (Munchen/München/Miinchen/Mynchen), grammatiska konstruktioner (Östersjön/ Östersjöns/Östersjö-*), gemen- och versaltillämpning (Malmö/MalmÖ/MALMÖ) och översättningar (Thames/Themsen), samt utesluta felstavningar, nonsensord (oftast OCR-misstag), missidentifierade personnamn (Bergvik) och ortnamn som uteslutande användes i en annan betydelse (Holmen). Resultatet blev en lista på 2171 geografiskt kodade entiteter. Bortfallet av termer i kodningsprocessen exemplifierar kombinationen av det geografiska språkbrukets dynamiska karaktär [19, s. 8–9], och den historiska OCR-avläsningens oundvikliga brus [20].

Sweriks korpus ligger också till grund för databasen riksdagsdebatter.se och med stöd av de kodade entiteterna kunde trender i förekommandet av olika grupper över tid kartläggas (t.ex. tyska platser). Nedan är ett exempel på hur en sådan kartläggning över perioden kan se ut.

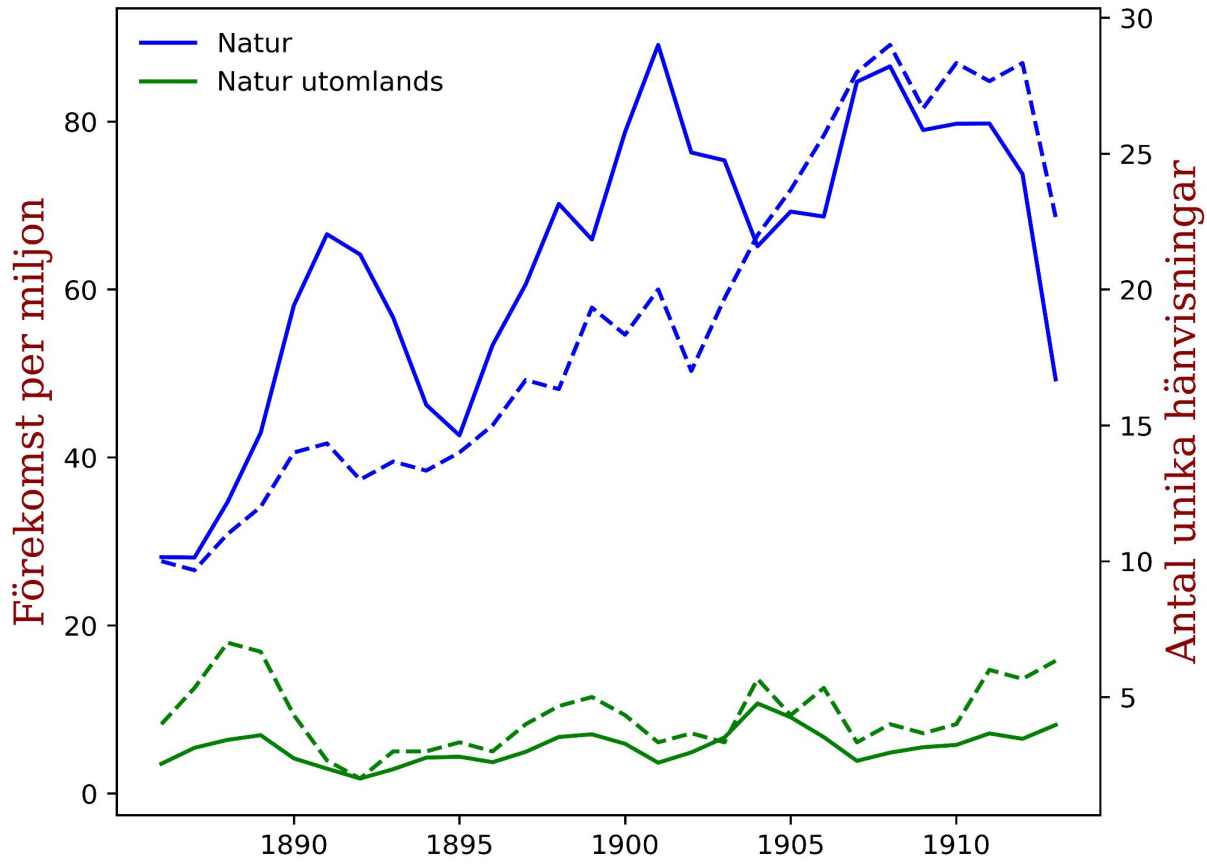


Figure 1: Exempel av kartläggning baserad på 242 platsnamn och områden som faller inom naturbeskrivningar i stil med sjöar, berg, naturområden, vikar, och så vidare. Hänvisningarna är uppdelade på inhemsk (blå) och utländsk (grön) natur. De heldragna linjerna läses via den vänstra y-

axeln och speglar den relativa, aggregerade frekvensen, beräknad som antalet träffar per en miljon tokens. De streckade linjerna läses via den högra y-axeln och speglar antalet unika hänvisningar för ett givet år. Visualiseringarna är baserad på ett rullande medelvärde över tre år.

Utifrån de identifierade trenderna drogs några grundläggande kvantitativa slutsatser. Dessa grundade sig främst i relativ ordfrekvens, men även antalet olika termer som förekom, samt kollokationer. De kvantitativa resonemangen fungerar som introduktioner till den mer substantiella, kvalitativa analysen av de relevanta texterna. Lejonparten av den analytiska insatsen utförs alltså i linje med traditionell närläsning, grundad i en aktiv (re)kontextualisering [21], i enlighet med en språkbrukshistorisk modell [22, s. 1–15].

Projektet resulterade i ett artikelutkast med namnet "Through the Eyes of the Germanic North: The International Geography of Swedish Parliamentary Debate 1887–1914" och har skickats till *Scandinavian Journal of History* för peer review. Vid jämförelse av de utländska referenserna blev det snabbt tydligt att Tyskland intog en särställning. Under perioden förekom 98 tyska platsnamn i debatterna, jämfört med 44 norska, 37 brittiska, 35 amerikanska (USA), 27 franska, och så vidare. De tyska platsnamnen har också den högsta sammantagna frekvensen genom hela perioden, med undantag för Norge några år innan sekelskiftet. En bidragande orsak är att landsnamnen uteslutits från jämförelserna då jag argumenterar för att dessa snarare används som politiska än geografiska entiteter. De tyska delstaterna driver alltså upp den tyska statistiken, men även i jämförelser mellan landsnamnen ligger "Tyskland" högt – om än något under "Norge".

Efter ytterligare granskning av resultaten stod det tydligt att bland de 25 vanligaste tyska platsnamnen så dominerade de nordtyska platsnamnen med 14 stycken, jämfört med 7 sydliga och 4 centrala. Den aggregerade frekvensen hos de nordtyska platsnamnen var också konsekvent nästan dubbelt så hög.

Utifrån denna kvantitativa utgångspunkt kunde sen kvalitativa läsningar belägga att Riksdagen hade en särskild relation till den tyska geografin under perioden och att jämförelserna med och beskrivningarna av de (nord-)tyska förhållandena var mer nyanserade och detaljerade än för andra länder. Kombinationen av språkkunskaper och i flera fall personliga erfarenheter gjorde riksdagsmännen väl förtrogna med tysk ekonomi och politik och den geografiska närheten bidrog till platsernas relevans. Tyska platser var centrala ekonomiska knytpunkter men hänvisades också till i som en del av reflektioner kring ländernas delade historia och som konkreta internationella exempel på politisk praktik – allt som oftast i termer av ett eftersträvansvärt ideal. Till den tidigare forskningen om Sveriges nära relation med Tyskland under perioden innan första världskriget lägger studien till en central aspekt i hur svensk politik internationaliserades genom sin relation till de tyska orterna. Hur riksdagspolitikernas kännedom om de tyska förhållandena tillät området att agera trygg och fast hållpunkt när man placerade Sverige i den europeiska kontexten. Fynden ger ytterligare och en mer mångfacetterad förståelse av den intima sammanflätningen av tysk och svensk politik vid det första världskrigets utbrott.

Martin Fridlund har med begreppet "digital historia 1.5" identifierat en typ av historiker som blandar traditionella och digitala metoder, använder sig av semi-automatiskt källurval och som är medveten om de digitala metodernas roll i forskningen men som inte själv kodar eller utvecklar modeller [23]. Beskrivningen stämmer på många sätt in på det här projektets tillvägagångssätt, förutom att min process resulterar i något som ligger mycket närmare traditionell idéhistoria än en balanserad kombination av kvantitativa och kvalitativa inslag. Man skulle kunna kalla mitt projekt en övning i digital historia 1.25. Det är baserat på digitala och semi-automatiska urvalsmetoder för att identifiera den relevanta primärlitteraturen, jag har dragit initiala och översiktliga kvantitativa slutsatser om materialets mest relevanta trender och jag har bearbetat den underliggande koden mer än en blackbox – men den slutgiltiga analysen har sin självklara tyngdpunkt i traditionella, kvalitativa metoder och uttryckssätt.

4. Några reflektioner

Tanken med BERT var aldrig att man ville ta fram en strömlinjeformad, användarvänlig blackbox-lösning och det märks såklart. BERTs höga prestationsförmåga inom en rad områden håller den relevant som ett stående inslag i datavetenskapliga publikationer om historiskt material [19]. Att så som beskrivits ovan hantera modellen mer eller mindre "naket" passar nog varken den digitala historikern

1.0 – som inte vill beblanda sig med koden bakom – eller 2.0 – som söker en mer beprövad, automatisk lösning där hallucinationer åtgärdats till den grad att resultaten kan stå för sig själva. Klumpigheten, vill jag dock hävda, bär med sig fördelen att en kvalitativ geografisk kodning är greppbar även för den digitalt skeptiske. I studier med begränsade digitala ambitioner har modellen kvalitativa styrkor både i och med sin transparens och förmågan att hantera den inneboende komplexiteten i historisk typonymy [24]. Det finns utrymme och anledning att diskutera listan med kodade entiteter på många vis. Man kan till exempel diskutera vad det innebär att den norska staden "Bergen" uteslutits från sökningarna då det inte går att särskilja från det svenska ordet. Det finns även flera geografiska namn, såsom "Boden" eller "Vaxholm", som används snarare synonymt med ett anläggningsnamn än som geografiska hänvisningar. Man skulle till och med kunna finna enstaka referenser som saknas från listan, men som urvalskriterium beror inte listans giltighet av de digitala metoder som hjälpt till att ta fram den; att NER-resultaten kodats manuellt gör att de kan bedömas och diskuteras som en kvalitativ produkt, oavsett den exakta vägen dit.

I slutändan väljs såklart forskningsverktyg på basis av faktorer såsom hur bra förklaringsvärde de har och hur exakta resultat de ger och där bedömer jag metoden ovan såsom alla andra. Jag vill dock slå ett slag för att den lätt klumpiga och förhållandevis analoga digitala idéhistorien kan kompensera för de många kvalitéer den saknar i förhållande till mer raffinerade lösningar. Utifrån de praktiska förutsättningarna som råder inom den vetenskapliga verksamheten så finns det fördelar i denna "manuella karaktär" och min förhoppning är att resultatet bör uppfattas som fullt kompatibel med traditionella idéhistoriska överväganden. Det finns på de flesta historiska institutioner helt enkelt mer expertis, erfarenhet och kommentarer att tillgå från kollegiet för en lista med geografiska namn än en kod. Detta är ingen radikal slutsats utan endast ännu ett konstaterande på att de historiska vetenskaperna vinner på ju mer bredd de kan visa upp, samtidigt som att kommunikation mellan de olika inriktningarna är avgörande för att maximera dessa vinster. I den traditionella och digitala historians spänningsfält finns en mängd positioner att inta och insikten från det här metodologiska greppet är att positioneringen också har ett medlande värde, som inte endast är *effort justification*.

5. Acknowledgements

Denna forskning har möjliggjorts av bidrag från Åke Wibergs stiftelse (H24-0183).

References

[1] W Said, Edward W. Orientalism. Vintage Books, New York, NY, 1979.
[2] Anderson, Benedict. Imagined Communities: Reflections on the Origin and Spread of Nationalism. Rev. ed., Verso, London, 1991.
[3] Götz, Norbert, and Janne Holmén. "Introduction to the Theme Issue: 'Mental Maps: Geographical and Historical Perspectives'." Journal of Cultural Geography 35.2 (2018): 157–61. doi:10.1080/08873631.2018.1426953.
[4] Benton, Lauren, "Spatial Histories of Empire." Itinerario XXX.3 (2006): 19–34.
[5] Schenk, Frithjof Benjamin. "Mental Maps. Die Konstruktion von geographischen Räumen in Europa seit der Aufklärung. "Geschichte und Gesellschaft 28.3 (2002): 493–514.
[6] Franzosi, Roberto. "Of Narrative Time and Space: Geography Meets History via Linguistics." Digital Scholarship in the Humanities 37.4 (2021): 982–96. doi:10.1093/llc/fqab090.
[7] Westerholt, René, Franz-Benjamin Mocnik, and Alexis Comber. "A Place for Place: Modelling and Analysing Platial Representations." Transactions in GIS 24.4 (2020): 811–18. doi:10.1111/tgis.12647.
[8] Szombara, Stanisław. "Using Different Mapping Techniques and GIS Programs in the Analysis and Visualisation of Mental Maps." Polish Cartographical Review 53.1 (2021): 91–104. doi:10.2478/pcr-2021-0008.
[9] K. Wagner, Kognitiver Raum: Orientierung – Mental Maps – Datenverwaltung, in: S. Günzel (Ed.), Raum. Ein interdisziplinäres Handbuch, J. B. Metzler, Stuttgart, 2010, pp. 234–49.

[10] Schneider, Ute. "Dimensions of Remapping: Heinrich Schiffers and His Mental Map of Africa." Journal of Cultural Geography 35.2 (2018): 162–88. doi:10.1080/08873631.2018.1426951.
[11] James, Laura M., Gamal Abdel Nasser, in: S. Casey and J. Wright (Eds.), Mental Maps in the Early Cold War Era, 1945–68, Basingstoke, Palgrave Macmillan, 2011, pp. 218–39.
[12] Varga, Mihai. "Mental Maps of Eastern Europe: States, Mentalities, Modernisation." Historical Sociological 35 (2022): 372–88.
[13] D. Mishkova, B. Trencsényi, Introduction, in: D. Mishkova and B. Trencsényi (Eds.), European Regions and Boundaries: A Conceptual History, Berghahn Books, Oxford, 2017, pp. 1–14.
[14] Guldi, Jo. "The Official Mind's View of Empire, in Miniature: Quantifying World Geography in Hansard's Parliamentary Debates." Journal of World History 32.2 (2021): 345–70. doi: 10.1353/jwh.2021.0028.
[15] J. L. Leidner and M.D. Lieberman. "Detecting Geographical References in the Form of Place Names and Associated Spatial Natural Language." SIGSPATIAL Special 3.2 (2011): 5–11, doi:10.1145/2047296.2047298.
[16] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, Attention Is All You Need, Advances in Neural Information Processing Systems 30 (2017). doi:10.48550/arXiv.1706.03762.
[17] KBLab, Huggingface, 2024. URL: https://huggingface.co/KBLab/bert-base-swedish-cased-ner.
[18] Swerik GitHub, 2025. URL: https://github.com/swerik-project/riksdagen-records?tab=readme-ov-file.
[19] M. Ehrmann, A. Hamdi, E.L. Pontes, M. Romanello, and A. Doucet, Named Entity Recognition and Classification in Historical Documents: A Survey, ACM Computing Surveys 56.2 (2023). doi:10.1145/3604931.
[20] M. Ehrmann, G. Colavizza, Y. Rochat, and F. Kaplan, "Diachronic Evaluation of NER Systems on Old Newspapers." Proceedings of the 13th Conference on Natural Language Processing (2016): 97–107.
[21] S. Bergwik, Omkontextualisering: Det kreativa bygget av sammanhang, in: S. Bergwik, L. Holmberg, and K. Dirke (Eds.), Konsten att kontextualisera: Om historisk förståelse och meningsskapande, Stockholm University Press, Stockholm, 2022, pp. 87–101.
[22] H. Björck, Folkhemsbyggare. Atlantis, Stockholm, 2008.
[23] M. Fridlund, Digital History 1.5: A Middle Way between Normal and Paradigmatic Digital Historical Research, in: M. Fridlund, M. Oiva, and P. Paju (Eds.), Digital Histories: Emergent Approaches within the New Digital History, Helsinki University Press, Helsinki, 2020, pp. 69–87.
[24] W. Zelinsky, On the Naming of Places and Kindred Things, in: S. J. Bronner (Ed.), Creativity and Tradition in Folklore: New Directions, Logan: Utah State University Press, 1992, pp. 179–184.

Shared Engagement in Digital Environments with Extended Reality and Tangible Interaction

Luis Quintero[1,*], Jordi Solsona[1], António Pinheiro Braga[1], Michael Björn[2], Uno Fors[1,*] and Harko Verhagen[1,*]

[1]Dept. of Computer and Systems Sciences (DSV), Stockholm University, 164 25, Stockholm, Sweden
[2]ConsumerLab, Ericsson Research, 164 40, Stockholm, Sweden

Abstract

Emergent interactive technologies – such as extended reality (XR) and its related subcategories augmented, virtual and mixed reality– are increasingly used in interdisciplinary research endeavors. These technologies aim to explore how smart glasses and headsets that overlay digital objects may support the design of collaborative experiences that enhance human interactions in the physical world. In this short paper, we briefly outline the possibilities of immersive technologies for research and how the Extrality Lab at Stockholm University serves as an infrastructure to prototype state-of-the-art solutions that merge physical tangible interaction and virtual environments in novel applications. We also describe how 3D digital tools may be used for research purposes, taking as an example the project SECE, which aims to study novel interactions, technology-supported artistic expressions, and the future of mobile computing in a cross-disciplinary team in Stockholm. More details about the Extrality Lab at https://extralitylab.dsv.su.se/.

Keywords

Virtual Reality, VR, Mixed Reality, MR, Extended Reality, XR, Human Computer Interaction, Interaction Design, Art, Performance, Collaboration, Telecommunication, 6G,

1. Introduction

The interplay between technologists and humanities scientists has enabled the advancement of scholarship in digital humanities, the interdisciplinary field concerned with the application of computational or digital methods to questions in humanities research, along with the critical examination of how digitalization shapes culture and society [1, 2]. Libraries and museums, which may be stereotypically attributed as analog settings, have been early adopters in the technological disruption. Even the first optical character recognition (OCR) algorithms were rapidly used for digitizing historical books at scales not possible with manual methods [3]. Similarly, more recent advances in 3D scanning and computer vision are the foundation for virtual tours of cultural heritage sites and creating digital replicas of archaeological artifacts and relics that require delicate physical preservation [4, 5]. In this paper, we discuss the possibilities of immersive digital media and tangible interaction as research methods in cross-disciplinary work. More specifically, we refer to the emergent technologies named Extended Reality (XR) and the potential of using the Extrality Lab at Stockholm University as a piece of infrastructure to support new research endeavors in digital humanities [6].

The acronym XR is an umbrella term for several tools that combine digital 3D objects with physical 3D space. It encompasses other technologies known as virtual, augmented, and mixed reality (VR/AR/MR, respectively). VR fully immerses users in computer-generated environments, while AR overlays digital information onto the real world, and MR allows digital and physical elements to interact in real time [7]. Researchers have largely studied the opportunities and challenges of XR experiences that are possible

Huminfra Conference 2025, Stockholm, 12–13 November 2025.
*Corresponding author.
luis.quintero@dsv.su.se (L. Quintero); jordi@dsv.su.se (J. Solsona); antonio.braga@dsv.su.se (A. Pinheiro Braga); michael.bjorn@ericsson.com (M. Björn); uno@dsv.su.se (U. Fors); verhagen@dsv.su.se (H. Verhagen)
0000-0002-6047-2793 (L. Quintero); 0000-0002-8951-6593 (J. Solsona); 0009-0005-0592-6798 (A. Pinheiro Braga); 0009-0000-6879-6410 (M. Björn); 0000-0002-3166-1640 (U. Fors); 0000-0002-7937-2944 (H. Verhagen)

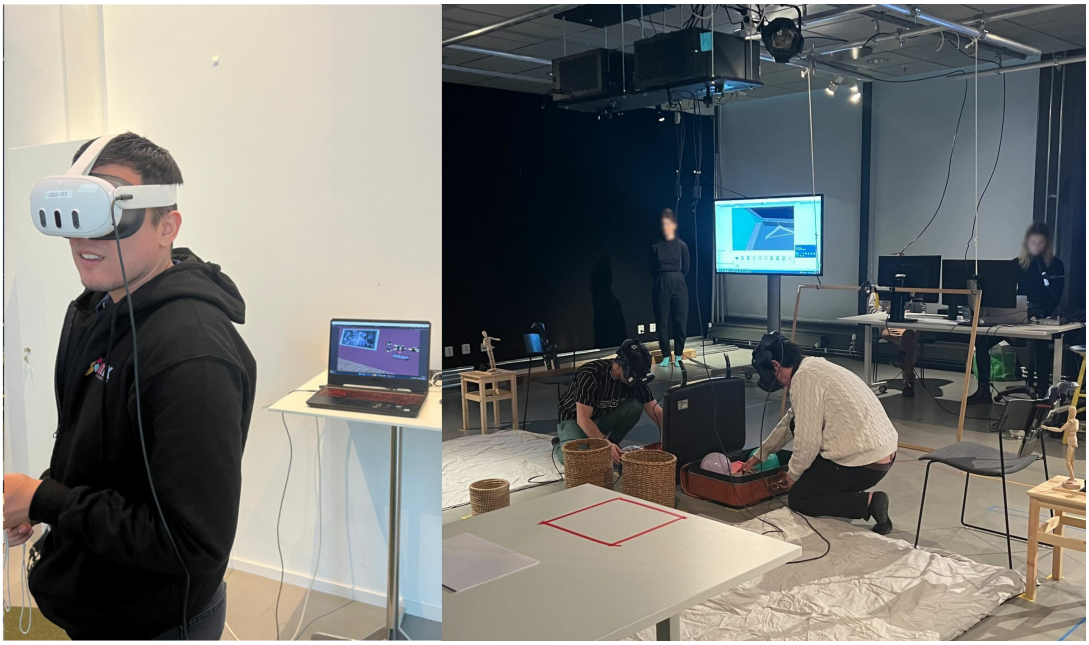


Figure 1: Resources at Extrality Lab: The multifunctional setup at Extrality Lab allows easy transition from standard seated settings to open spaces for public demonstrations.*Left*: A person wearing a headset to explore a single-user XR experience. *Right:* Two participants engage in a shared narrative that combines virtual worlds and physical objects with tangible interaction.

beyond purely physical or digital spaces. More recently, policy makers have also outlined initiatives to explore XR, such as the European roadmap for education and healthcare [8] or the Swedish funding in XR for sustainable innovation[1].

An underlying principle of these immersive systems is to create *hands-free* digital applications that let users collaborate without diverting attention to physical screens, including phones. The main interface in XR utilizes smartglasses and headsets as a medium to interact with 3D content through natural gestures (e.g., voice, hands, eye movements) [9, 10]. For instance, a museum visitor wanting to know more about a piece of art would get digital information projected directly onto their glasses, rather than reading a label or a screen [11]. Similarly, a collaborative project with XR would let people maintain interactions while standing and facing each other, where the relevant information appears seamlessly as virtual panels in the physical room rather than having to sit in front of physical screens [12, 13].

Overall, XR may support digital humanities by creating immersive, interactive environments that allow scholars and the public to experience cultural, historical, and artistic artifacts in new ways. For researchers, it enables novel forms of data visualization, exploration of sensitive objects, or interactions blending storytelling with embodied experiences that are aware of the physical space. In the remainder of the paper, we describe Extrality Lab as a physical space that may be used as an infrastructure for the exploration of XR for varied use cases, and also summarize the project SECE as a successful example that leverages the infrastructure to explore questions related to interaction design, future 6G networks, and technology-supported artistic performances in other words blending research from computer science, engineering, and the arts.

2. Extrality Lab: Extended and Tangible Reality Lab

The Extrality Lab is a central research environment dedicated to advancing knowledge in immersive systems, tangible interaction, and embodied computing, as well as enabling novel digital humanities research approaches. It may be considered a suitable infrastructure for digital humanities [2]. Equipped

[1]https://www.vinnova.se/en/calls-for-proposals/emerging-technology-solutions/feasibility-studies-xr-sweden-6g-2024-01699/

with a full suite of XR devices, professional media production tools, and a fabrication workshop for physical prototyping, the lab provides researchers with the means to conduct systematic experimentation on the design and evaluation of emerging interaction paradigms.

The physical and human resources available at Extrality Lab have served as a scientific and technical framework for existing research projects at higher education institutions and through collaborative partnerships. Some of the developed immersive experiences have targeted the design and evaluation of educational tools to explain abstract concepts, such as learning introductory-level programming structures [14] or encryption algorithms for cybersecurity [15]. Another line of research in interaction design and user modeling has explored how novel 3D interfaces may affect human factors like the sense of embodiment toward digital avatars [16] or the elicited emotions from virtual content [17].

Lastly, the intersection of XR and related immersive technologies with artistic performances has facilitated the understanding of how performers and audience members interact in worlds with virtual and tangible interaction (see Figure 1). These activities explore novel opportunities of participation, their collaborative roles, and their agency within artistic productions [18]. The infrastructure at the Extrality Lab has been available to keep exploring similar questions through the lens of emergent technologies, such as in the SECE project described below, which represents the future of mobile computing and shared engagement in immersive digital environments.

3. SECE: Shared Engagement in Cultural Events with Mixed Reality

3.1. Motivation

MR in the context of multi-user experiences is relatively unexplored due to technological restrictions that were overcome only in 2024, allowing headsets to run colocated applications in a wireless mode without depending on desktop computers for 3D rendering. Therefore, SECE is one of the first projects to build MR collaborative experiences for shared engagement in digital environments and test them in a real-life outdoor setting. The project is a collaboration between the Extrality Lab at Stockholm University, the ConsumerLab at Ericsson Research, and Kulturhuset Stadsteatern. This project explores multidisciplinary research questions related to XR interactions, telecommunications (5/6G, WiFi 6e and beyond), and artistic performances, such as: *how can immersive MR interactions enable new dynamics between actors and audience in an interactive performance?*, *what are the network challenges when running a mobile MR experience outdoors?*, and *how can MR support novel and meaningful social interactions when designing interactive artistic performances?*.

Previous XR research has primarily addressed either single-user experiences or collaborative systems relying on multi-device configurations. Current headsets feature real-time see-through capabilities with co-location features through the real-time matching of a room's point cloud, allowing several users placed in the same space to interact with the same aligned virtual content.

As for use of VR in both learning and the arts, as presented at HumInfra Conference 2024[19], new technology brings opportunities but also challenges on how to organize the process of synchronizing the work of all parties involved. XR comes with its own challenges that need addressing when designing and implementing. For SECE, the core is the novel multi-user colocated application that allows transitions between reality, MR, and VR using digital portals to support the study of how smartglasses and headsets may work as a platform to perform synchronous and collaborative tasks that seamlessly transition between fully digital and fully real settings (see Figure 2, and next section). A full description of the technical setup has been published previously in a scientific conference [20, 21], and more related information about SECE is available at the project's website[2].

[2]https://www.su.se/english/research/research-projects/mixed-reality-shared-engagement-in-cultural-events-sece

Figure 2: Snapshot of the XR colocated experience SECE: *Left*: Several participants wearing XR headsets interact with co-located virtual cubes (blue) to design an art performance combining real and virtual objects. *Right*: Perspective of a participant not wearing XR headset and therefore unable to see the digital content in the physical space.

3.2. Demo and Implementation

XR Interaction Applications: Previous work on multi-user XR systems for cultural heritage found that object interactivity and user-generated content significantly enhance engagement [22]. Therefore, in the SECE project, we developed applications to facilitate the design of artistic participatory experiences. First, the leader of the artistic performance can perform *object placement* (see Figure 2-Left), where a set of virtual 3D elements will float in the physical space and become interactable for anyone wearing a headset. A second application enables *mid-air drawing*, where the virtual pens can be placed in the 3D space and the participants can create scribbles and traces in the 3D space. The third application supports *character control*, where the movement trajectory of a character, a butterfly in the SECE case, is controlled by the users to move the virtual elements in the physical room. Lastly, spatial audio sources can be placed to complement a multisensory experience. All virtual elements are updated in real-time to give all participants a synchronous immersive experience.

Performance Preparation: The first participant entering the XR experience acts as the session's leader and should set up a digital scan of the physical room to be processed in the headset. This process captures point-cloud data from the physical space and the contextual information of the surrounding physical objects, such as tables, doors, or plants. The room's contextual data creates the spatial anchor, enabling the colocation of other participants. When the other participants access the same application in their headsets, the room data is synchronized to let everyone see the virtual objects in the same place, despite looking at them from different perspectives.

Artistic Storyline: The leader of the artistic performance can use the available XR interaction applications to control the narrative of the immersive experience. The preliminary performative storyline of SECE unfolds in three acts.

Evaluation: The objectives of the SECE project involve analysing how XR reshapes sensemaking, artistic expression, and audience engagement in collaborative performances. The first analysis moment was conducted indoors at Kulturhuset in June 2025, and the final outdoor performance is planned for June 2026 at Sergelstorg. The data collected is primarily qualitative, with interviews and video recordings that inform the possibilities for the future design of mobile digital experiences with immersive systems, specifically in artistic settings.

4. Conclusion

This paper describes Extended Reality (XR) as a tool or method that may support current research in digital humanities. More specifically, we describe how the Extrality Lab can serve as an infrastructure for the design and development of such projects at the Department of Computer and Systems Sciences (DSV) of Stockholm University. Lastly, we present the SECE project as a case of multidisciplinary work that creates a complex MR interactive experience to address questions related to interaction design, telecommunications, and performative arts.

Acknowledgments

The authors would like to acknowledge all team members of the research project "Mixed Reality Shared Engagement in Cultural Events (SECE)" at Kulturhuset Stadsteatern, Stockholm University, and Ericsson Research. This work is funded by DSV at Stockholm University and Digital Futures (see https://www.digitalfutures.kth.se/).

References

[1] C. Warwick, M. Terras, J. Nyhan, Digital Humanities in Practice, Facet, 2012, p. xiii–xx. doi:10.29085/9781856049054.

[2] D. M. Berry, A. Fagerjord, Digital Humanities, Polity, 2017.

[3] M. J. Wachowiak, B. V. Karas, 3d scanning and replication for museum and cultural heritage applications, Journal of the American Institute for Conservation 48 (2009) 141–158. doi:10.1179/019713609804516992.

[4] H. Wang, Z. Gao, X. Zhang, J. Du, Y. Xu, Z. Wang, Gamifying cultural heritage: Exploring the potential of immersive virtual exhibitions, Telematics and Informatics Reports 15 (2024) 100150. doi:10.1016/j.teler.2024.100150.

[5] S. Robson, S. MacDonald, G. Were, M. Hess, 3D recording and museums, Facet, 2012, p. 91–116.

[6] M. S. Anwar, J. Yang, J. Frnda, A. Choi, N. Baghaei, M. Ali, Metaverse and xr for cultural heritage education: applications, standards, architecture, and technological insights for enhanced immersive experience, Virtual Reality 29 (2025). doi:10.1007/s10055-025-01126-z.

[7] P. Milgram, F. Kishino, A Taxonomy of Mixed Reality Visual Displays, IEICE Trans. Information Systems E77-D, no. 12 (1994) 1321–1329.

[8] European Commission, Directorate-General for Communications Networks, Content and Technology, C. Boel, K. Dekeyser, F. Depaepe, L. Quintero, T. Daele, B. Wiederhold, Extended reality : opportunities, success stories and challenges (health, education), Publications Office of the European Union, 2023. doi:10.2759/121671.

[9] P. Manakhov, L. Sidenmark, K. Pfeuffer, H. Gellersen, Gaze on the Go: Effect of Spatial Reference Frame on Visual Target Acquisition During Physical Locomotion in Extended Reality, in: CHI '24 : Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems, Association for Computing Machinery, 2024, pp. 1–16. doi:10.1145/3613904.3642915.

[10] K. Pfeuffer, B. Mayer, D. Mardanbegi, H. Gellersen, Gaze + pinch interaction in virtual reality, in: Proceedings of the 5th Symposium on Spatial User Interaction, ACM, Brighton United Kingdom, 2017, pp. 99–108. doi:10.1145/3131277.3132180.

[11] Y. Shu, Virtual tours in digital museums: Vr and ar technologies in the modern cultural space, International Journal of Human–Computer Interaction (2025) 1–17. doi:10.1080/10447318.2025.2537799.

[12] M. K. Bekele, R. Pierdicca, E. Frontoni, E. S. Malinverni, J. Gain, A survey of augmented, virtual, and mixed reality for cultural heritage, Journal on Computing and Cultural Heritage 11 (2018) 1–36. doi:10.1145/3145534.

[13] Y. Li, R. Yang, J. Zou, H. Xu, F. Tian, Human-centric virtual museum: Redefining the museum experience through immersive and interactive environments, International Journal of Human–Computer Interaction 41 (2024) 8426–8437. doi:10.1080/10447318.2024.2408861.

[14] J. Ekman, J. Solsona, L. Quintero, Codeseum: Learning Introductory Programming Concepts through Virtual Reality Puzzles, in: ACM International Conference on Interactive Media Experiences, ACM, 2024, pp. 192–200. doi:10.1145/3639701.3656306.

[15] M. Bernsland, A. Moshfegh, K. Lindén, S. Bajin, L. Quintero, J. Solsona Belenguer, A. Rostami, CS:NO – an Extended Reality Experience for Cyber Security Education, in: ACM International Conference on Interactive Media Experiences, Aveiro, Portugal, 2022, pp. 287–292. doi:10.1145/3505284.3532971, series Title: IMX '22.

[16] J. Ulrichs, A. Matviienko, L. Quintero, Effects of Third-Person Locomotion Techniques on Sense of Embodiment in Virtual Reality, in: Proceedings of the International Conference on Mobile and Ubiquitous Multimedia, MUM '24, Association for Computing Machinery, New York, NY, USA, 2024, pp. 72–81. doi:10.1145/3701571.3701598.

[17] L. Quintero, User Modeling for Adaptive Virtual Reality Experiences: Personalization from Behavioral and Physiological Time Series, Department of Computer and Systems Sciences, Stockholm University, Stockholm, 2023.

[18] A. Rostami, Interweaving Technology: Understanding the Design and Experience of Interactive Performances, Department of Computer and Systems Sciences, Stockholm University, Stockholm, 2020.

[19] HiC-2024 Abstract Submissions, Göteborg Universitet, 2024. URL: https://ecp.ep.liu.se/index.php/hic/issue/view/84/88.

[20] L. Quintero, A. M. B. M. Pinheiro Braga, N. Petersson, U. G. Fors, Transitional Portals for Participatory Co-Located Cross-Reality Experiences, in: Proceedings of the 2025 ACM International Conference on Interactive Media Experiences, IMX '25, Association for Computing Machinery, New York, NY, USA, 2025, pp. 368–371. doi:10.1145/3706370.3731708.

[21] L. Quintero, E. Bennaceur, L. Ahlnäs, M. Bjorn, Hands-On Orchestra: Hand-based Interactive Manipulation of Spatial 3D Audio in Mixed Reality, in: Proceedings of the 2025 ACM International Conference on Interactive Media Experiences, IMX '25, Association for Computing Machinery, New York, NY, USA, 2025, pp. 342–345. doi:10.1145/3706370.3731713.

[22] Y. Li, E. Ch'ng, S. Cobb, Factors Influencing Engagement in Hybrid Virtual and Augmented Reality, ACM Trans. Comput.-Hum. Interact. 30 (2023) 65:1–65:27. doi:10.1145/3589952.

Exploring Patient Organization Periodicals with the Topic Timelines Text Visualization Method

Maria Skeppstedt[1,*], Adam Maen[2], Vera Danilova[3], Gijs Aangenendt[3], Andrew Burchell[3] and Ylva Söderfeldt[3]

[1]*Stockholm University, Department of Linguistics, Stockholm, Sweden*
[2]*Uppsala University, Centre for Digital Humanities and Social Sciences, Department of ALM, Uppsala, Sweden*
[3]*Uppsala University, Department of History of Science and Ideas, Uppsala, Sweden*

Abstract

The text visualization technique Topic Timelines offers a compact visualization to represent the evolution and clustering of topics over time, while also providing direct access to the texts in which these topics appear. In this paper, we describe how Topic Timelines was further developed within the ActDisease project, by adding functionality for generating timelines using different types of topic extraction techniques and connecting the visualization to existing interfaces for the close reading of texts. Additionally, we evaluate how the updated temporal topic overview can support corpus exploration.

The experiments were conducted on a digitalized corpus from the ActDisease project, consisting of patient organization periodicals from the Swedish Diabetes Association, published between 1949 and 1990. Timelines were generated based on topics extracted using sentence transformers clustering and integrated with the ActDisease text database interface – a user interface developed for exploring and reading texts digitalized within the project.

Keywords

Text visualization, Automatic topic extraction, Sentence transformers, Patient organizations

1. Introduction

The ActDisease project studies the history of patient organizations in twentieth century Europe, combining traditional modes of analysis with computer-based methods[1]. Periodicals from selected British, French, German and Swedish patient organizations have been digitalized [1] and made searchable through a graphical user interface developed within the project. The interface enables the historians within the project to carry out keyword searches, filter on type of publication and language, as well as browse the scanned magazine issues page by page (see Figures 1 and 2).

In addition to searching and browsing the materials using the interface, the Topic Timelines visualization technique has been developed within the project as an additional method for exploring the patient organization periodicals [2]. By visualizing the evolution of topics within a corpus over time, an overview of the corpus content is provided. However, previous techniques for visualizing topics in temporal text collections [3, 4, 5, 6, 7, 8, 9, 10, 11, 12] typically aim to support quantitative data exploration, and therefore only provide a temporal overview in an aggregated format, without any direct connection to the texts from which the data is extracted. Topic Timelines, in contrast, is specifically designed to support historical research, offering a compact and information-rich overview visualization of topics in a corpus, while also providing direct access to the texts in which these topics appear[2].

Huminfra Conference 2025, Stockholm, 12–13 November 2025.
*Corresponding author.
maria.skeppstedt@ling.su.se (M. Skeppstedt); adam.maen@abm.uu.se (A. Maen); vera.danilova@idehist.uu.se (V. Danilova); gijs.aangenendt@idehist.uu.se (G. Aangenendt); andrew.burchell@idehist.uu.se (A. Burchell); ylva.soderfeldt@idehist.uu.se (Y. Söderfeldt)

[1]https://www.actdisease.org
[2]The code for creating the visualization, as well as examples of how the visualization can be used is found here: https://github.com/CDHUppsala/topic-timelines

This paper discusses the most recent developments of Topic Timelines, focusing specifically on the added functionality for generating timelines using different types of topic extraction techniques as well as the possibility to use the visualization as a method for navigation in existing database interfaces. As a case study for testing and evaluating the updated Topic Timelines, we generated a visualization for one of the ActDisease corpora – periodicals published by the Swedish Diabetes Association between the years 1949 and 1990[3] – and connected it to the graphical user interface for the ActDisease database.

2. Adapting and connecting Topic Timelines

In previous versions [2, 13], Topic Timelines visualizations could only be generated from the output of one specific topic extraction tool, the Topics2Themes tool [14]. To make it possible to apply the visualization technique to different types of topic extraction methods, we updated the Topic Timelines Python module, enabling it to create visualizations based on data provided in a tabular format[4]. There are many different topic extraction methods, each of them with different advantages. The previously used Topics2Themes tool extracts topics based on word co-occurrence patterns, and thereby provides topics that are relatively transparent to the user. In the case study described below, we instead used topic extraction based on clustering of sentence transformers, a technique that, e.g., allows the user to extract topics from multi-lingual corpora.

In addition, we explored the possibility to extend the support for organizing the topics on the timeline. Analyzing and comparing many topics on a timeline is difficult, even on a large computer screen. The Topic Timelines implementation, therefore, provides the user with the possibility to manually determine the order in which the extracted topics appear. We have previously used this functionality to manually group topics that share the same characteristics or themes. In this iteration of Topic Timelines development, we experimented with using topic clustering to instead automatically generate such high-level topic groups.

The third focus of this study was an integration of the timeline visualization with the graphical user interface for the ActDisease database, to investigate whether a thematic exploration of the database contents could complement the traditional ways of searching and browsing. The Topic Timelines Python module makes it possible to link each text visualized to a URL, which can be reached by clicking in the graph. We here linked each text to a corresponding URL in the ActDisease database interface. To make this connection, there must be a naming consistency between the URLs in the graphical user interface and the file names in the text corpus from which topics are extracted. We, therefore, updated the ActDisease interface so it uses "Clean URLs", where the URL address matches the filename in the text corpus. The user can thereby start their exploration by identifying potentially interesting texts in the graph, based on the timeline visualization, and then click on these texts to reach the standard ActDisease database interface. In the database interface, users can then inspect the scanned image and the extracted text, but also browse adjacent pages of the selected text. This is a development compared to previous studies [2, 13], where the user was directed to a separate webpage displaying only the scanned image, without any links to adjacent pages in the magazine.

3. Case Study

To test and evaluate the developments of Topic Timelines, we generated a visualization for the corpus of periodicals published by the Swedish Diabetes Association and connected it with the graphical user interface for the ActDisease database. Finally, the historians in the project inspected the timelines and the automatically generated ordering of the topics.

[3]The materials scanned by Gothenburg University Library can be found here: https://gupea.ub.gu.se/handle/2077/64597
[4]An example of this format can be found here: https://github.com/CDHUppsala/topic-timelines/tree/main/diabetes_topics

3.1. Method and configuration parameters for the topic extraction

The method used for creating topics was sentence transformers [15], which is a technique for encoding the meaning of a text unit into a vector. The distance between these vectors corresponds to the similarity between texts, and it is therefore possible to extract topics by performing a distance-based clustering of the vectors. Clusters of vectors then correspond to groups of texts belonging to the same topic.

We encoded the text units (i.e., in most cases the pages)[5], into vectors using the *paraphrase-multilingual-MiniLM-L12-v2*[6] SentenceTransformer [15]. We then applied the functionality for agglomerative clustering, which is available through scikit-learn [16], on these vectors. Agglomerative clustering is a clustering method that starts with each vector belonging to its own separate cluster, and then iteratively merges the two clusters that are closest to each other into larger clusters, until a user-specified distance threshold is met.

As stressed by Da [17, p 625], the output from automatic topic extraction is heavily dependent on the methods used and the parameters settings selected. Therefore, the topics extracted and visualized should be looked upon as a set of potentially interesting topics in the corpus, rather than as *the* topics in the corpus[7]. When experimenting with different configurations settings, our aim was consequently to render timelines containing topics with the potential to be interesting to the user. Therefore, to generate topics that were focused and specific enough, we varied the threshold for when to stop the merge of clusters. In addition, to avoid clusters containing too few elements, we implemented the functionality of allowing small clusters to be merged with the closest neighbouring cluster, regardless of distance. Finally, to avoid topics being created based on similarity between uninteresting function words, rather than between content words, we applied stop word filtering before encoding the texts into vectors.

We iteratively refined the clusters, resulting in the following configuration: a cosine distance threshold of 0.47, a minimum cluster size of 50 texts, and the standard Swedish stop word list from NLTK [19], extended with 80 corpus-specific stop words. This resulted in 68 topics being extracted and visualized.

Also for creating the high-level clusters, agglomerative clustering was used, with a cosine distance threshold value of 0.35. This resulted in 12 high-level groups, and there were 7 outlier topics, not fitting in any of the groups.

3.2. The resulting visualization

The resulting timeline graph is shown in Figure 3, and a description of the graph components is shown in Figure 4. Each one of the 68 topics is represented by a horizontal lane, and each text in the corpus is represented by a vertical line[8]. The position on the timeline for the vertical text-line is determined by the date when the text was published. In the case of the periodical issues, several texts are published on the same date. As a solution, the text line corresponding to the first page is placed on the position corresponding to the publication date, and the lines for the following pages are moved slightly to the right. The automatic division into high-level groups of topics is represented by color coding.

When a text belongs to a topic, this is indicated by a circle in the graph at the point where the text line and topic lane intersect. The more typical a text for a topic, the larger the circle representing the text-topic connection[9]. The circles are semi-transparent, which has the effect that the prevalence of the same topic in many adjacent texts is represented by a pattern of overlapping, semi-transparent circles.

[5]The corpus investigated is divided into pages. For most cases, a text unit therefore means a page. However, long pages had to be split into smaller subtext units, to accommodate for the maximum token limit of the sentence transformer used (512 tokens). For these pages, blank lines – which often signal paragraph breaks – were used for splitting the texts. The 1 112 pages in the corpus that contained more than 300 words were split into subpages, resulting in a total of 9 775 texts to cluster.

[6]https://huggingface.co/DataikuNLP/paraphrase-multilingual-MiniLM-L12-v2

[7]As a more objective measure of the corpus content, we would argue that frequency-based methods instead form a better choice [18].

[8]We refer to https://github.com/CDHUppsala/topic-timelines and previous publications for a more detailed description of the visualization technique.

[9]We here measured how typical the text was, by its distance to the centroid of the cluster. I.e., the closer to the centroid of the cluster, the more typical is the text.

When clicking on a circle, the user is directed to the ActDisease database interface, showing the original text in which the topic occurs (as shown in Figure 5).

3.3. Reading the timelines

The most distant view of the graph is provided by the color-coded grouping into high-level topics. This view tells us that the corpus contains topics related to (a) nutrition, (b) food and sweeteners, (c) insulin and measuring of blood sugar, (d) insulin pumps and medical effects of diabetes, (e) insulin, activities for members, (f) children and youth, (g) members and organizational matters, (h) disability and economical matters, (i) the organization's chairperson and (perhaps) different types of letters, (j) travel, (k) economical matters and taking insulin, and (l) (perhaps) personal patient stories.

By zooming in one level into the visualization, we can move from the high-level clusters to study the temporal characteristics of the individual topics. For instance, topic 25 (about insulin) occurs frequently during the entire time period studied, topic 22 (about economical matters) occurs regularly once per year during a large part of the time series, the topics on patient stories become frequent during the 1980s, and finally topic 40 (about driving licenses), is very concentrated to the year 1965. This indicates that certain themes emerge at "flashpoints" – in the case of topic 40, campaigning around medical certification for driving – while others represent the consistency of organizational administration (financial reporting) or more slow-progressing cultural changes within the formatting and editorial choices of the magazine (the foregrounding of personal narratives and experiences). Advertisements likewise evolve differently, e.g., while topics 14 (sweeteners) and 53 (self-testing kits for urine) cover a long chronological range, the marketing of newer products which only entered the market during the 1980s (topics 46 and 61) is rendered more pronounced by the visualization.

Finally, the most detailed level is accessed by clicking on the circles to reach the original texts, as they are presented in the ActDisease database interface. For instance, a closer study of the texts related to the "advertisement" topics, allows us to see that the high concentration of these topics in the timeline visualization in the 1980s is not only a testament to the substantial advertising campaigns led by manufacturers (as shown in the example in Figure 5) but is also a consequence of, e.g., technical devices appearing in other parts of the periodical, including reviews and personal narratives of use.

3.4. Discussion

The principal benefit of the integration between the database and the timeline is that it simplifies not only navigation between the data points but also contextualization of the results. This allows the automatic topic extraction to exist in dialogue with more empirical methodologies, as we preserve the ability to move backwards and forwards in the magazine from a relevant data point as well as the physical experience of browsing the magazine as an artefact. This movement between the different "levels" of the corpora – as statistical data and modelled topics on the one hand, and textual, visual and formatted documents (albeit mediated through a digitalized database) on the other – offers the potential for fruitful and more intuitive uses by historians whose primary training is not in quantitative methods. It also allows users to merge more traditional and digital methodologies by collapsing the digital/material divide between the sources.

It can further be noted that the automated generation of topic groups suggests an interesting tension in the function of these visualizations. Here, the (auto)generated high-level clusters suggest broader themes which are of interest to historians: the mediation of consumption and self-care practices by people with diabetes (clusters a to d, h and k), the administrative life of the patient organizations themselves (clusters e, f, g, h and i), and the rise of individual narrative practices in the later half of the twentieth century (potentially clusters i and l). Yet cutting across the higher-level clusters are topics which also offer entry points into the different genres of text existing in the magazine corpora. Topics 14, 46, 53 and 61 are advertisements for a range of specialized products for people with diabetes, including sugar-free sweeteners (14), urine glucose-testing kits (53), and digital blood-glucose monitors (46 and 61). In previous iterations of these timelines, historians have sometimes tended to classify as much by genre as by theme – to make these advertisements more visible. As this case demonstrates, there are no "correct"

or “incorrect” classifications and clusters, and it is helpful to consider one’s own clustering in the context of other possible classification logics.

Finally, based on what topics were extracted, it seems that the topic timeline method is unusually suited to certain kinds of textual genres – and especially to the advertisements – because they are relatively stable texts in temporal and historical terms. Brand names, slogans, key selling points, and product designations often remain unchanged across long time periods, even if the advertisements themselves could be renewed or updated regularly to maintain reader interest. This is visible from the preponderance of proprietary names such as “sionon” (topic 14), “reflolux” and “boehringer [mannheim]” (topic 46), or “clinitest” and “ames” (topic 53) in the topic timeline. It is possible that this relative textual stability produces a slight bias within the generation of topic timelines. In addition, the format of the magazines, and the marketing economics of the manufacturing companies, meant that many of the advertisements for the testing devices were full-page by the 1980s (for an example, see Figure 5). Since advertisements are relatively short, and the texts were normally split for clustering on page level, this might result in a higher degree of coherency for the full-page advertisements than for the other texts, which might also have introduced a bias for advertisements.

4. Conclusions and future work

Overall, we would suggest that interactive navigation of the magazine database by automatic topic extraction and timeline visualizations offers potential for historians beyond usual keyword searches. Such methods can allow historians to pass between the “levels” of distant versus close textual reading, while also preserving the ability to interact with the visual and aesthetic dimensions of the magazines. Nonetheless, it is preferable to retain the flexibility with regard to the high-level clustering of topics: automated versions of this can prove useful in stimulating fresh perspectives, but they may have limitations when it comes to the presentation of results.

As a next step, we will develop the sentence transformers clustering described here into a Python module generally applicable on temporal corpora. We then aim to make the text clustering and the Topic Timelines visualization available as a resource within the Huminfra research infrastructure.

We also plan to employ the genre classifier [20] developed for the corpora digitalized within the ActDisease project to more closely investigate the relationship between genre and topic. For instance, by creating topics only for texts belonging to a certain text genre, or by investigating the prevalence of different genres within automatically extracted topics. This work also includes an exploration of more semantically informed methods for dividing the pages into subtexts.

Additionally, we are investigating further integration of the database into the visualization. This would enable the creation of an interface that could allow the user to browse the database not just chronologically by material in the same issue/volume of the magazine, but also according to related data points in the topic modeling and timeline. This would open up new possibilities for reading the database against the chronological and thematic grain, while preserving the possibility to still explore the magazines in a more traditional way.

Acknowledgements

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

The development of Topic Timelines has been partly supported by Huminfra (the Swedish Research Council, grant 2021-00176) and InfraVis (the Swedish Research Council, grant 2021-00181).

References

[1] G. Aangenendt, M. Skeppstedt, Y. Söderfeldt, Curating a historical source corpus of 20th century patient organization periodicals, in: Proceedings of the Huminfra Conference (HiC 2024), 2024, pp. 76–82. URL: https://ecp.ep.liu.se/index.php/hic/article/view/895. doi:10.3384/ecp205011.

[2] Y. Söderfeldt, A. Burchell, J. Reed, M. Skeppstedt, Topic timelines for enabling close and distant reading of discursive shifts. a pilot case using periodicals of european diabetes organizations, Journal of Open Humanities Data (2025). doi:10.5334/johd.286.

[3] M. Grootendorst, Dynamic topic modeling, visualization, https://maartengr.github.io/BERTopic/getting_started/topicsovertime/topicsovertime.html, 2023.

[4] D. Blei, J. Lafferty, Dynamic topic models, in: ACM International Conference Proceeding Series; Vol. 148: Proceedings of the 23rd international conference on Machine learning; 25-29 June 2006, ACM, 2006, pp. 113–120.

[5] S. Sheehan, S. Luz, M. Masoodian, TeMoTopic: Temporal mosaic visualisation of topic distribution, keywords, and context, in: H. Toivonen, M. Boggia (Eds.), Proceedings of the EACL Hackashop on News Media Content Analysis and Automated Report Generation, Association for Computational Linguistics, Online, 2021, pp. 56–61. URL: https://aclanthology.org/2021.hackashop-1.8.

[6] S. Malik, A. Smith, T. Hawes, P. Papadatos, J. Li, C. Dunne, B. Shneiderman, Topicflow: visualizing topic alignment of twitter data over time, in: Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ASONAM '13, Association for Computing Machinery, New York, NY, USA, 2013, p. 720–726. URL: https://doi.org/10.1145/2492517.2492639. doi:10.1145/2492517.2492639.

[7] C. Meaney, M. Escobar, T. A. Stukel, P. C. Austin, L. Jaakkimainen, Comparison of methods for estimating temporal topic models from primary care clinical text data: Retrospective closed cohort study, JMIR medical informatics (2022).

[8] S. Gad, W. Javed, S. Ghani, N. Elmqvist, T. Ewing, K. N. Hampton, N. Ramakrishnan, Themedelta: Dynamic segmentations over temporal topic models, IEEE Transactions on Visualization and Computer Graphics 21 (2015) 672–685. doi:10.1109/TVCG.2014.2388208.

[9] S. Havre, B. Hetzler, L. Nowell, Themeriver: visualizing theme changes over time, in: IEEE Symposium on Information Visualization 2000. INFOVIS 2000. Proceedings, 2000, pp. 115–123. doi:10.1109/INFVIS.2000.885098.

[10] N. Günnemann, M. Derntl, R. Klamma, M. Jarke, An interactive system for visual analytics of dynamic topic models, Datenbank-Spektrum 13 (2013) 213–223. URL: https://doi.org/10.1007/s13222-013-0134-x. doi:10.1007/s13222-013-0134-x.

[11] N. Günnemann, D-vita: A visual interactive text analysis system using dynamic topic mining, in: Datenbanksysteme für Business, Technologie und Web, 2013. URL: https://api.semanticscholar.org/CorpusID:15848321.

[12] W. Cui, S. Liu, L. Tan, C. Shi, Y. Song, Z. Gao, H. Qu, X. Tong, Textflow: Towards better understanding of evolving topics in text, IEEE Trans. Vis. Comput. Graph. 17 (2011) 2412–2421. URL: https://doi.org/10.1109/TVCG.2011.239. doi:10.1109/TVCG.2011.239.

[13] M. Skeppstedt, G. Aangenendt, V. Danilova, Y. Söderfeldt, Topics in periodicals from the swedish diabetes association 1949 – 1990: Extending the topic modelling tool topics2themes with a timeline visualisation, in: Selected papers from the CLARIN Annual Conference 2023, Linköping University Electronic Press, 2024. doi:https://doi.org/10.3384/ecp210015.

[14] M. Skeppstedt, K. Kucher, M. Stede, A. Kerren, Topics2Themes: Computer-assisted argument extraction by visual analysis of important topics, in: Proceedings of the LREC Workshop on Visualization as Added Value in the Development, Use and Evaluation of Language Resources, 2018, pp. 9–16.

[15] N. Reimers, I. Gurevych, Sentence-bert: Sentence embeddings using siamese bert-networks, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing,

Association for Computational Linguistics, 2019. URL: http://arxiv.org/abs/1908.10084.
[16] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay, Scikit-learn: Machine learning in Python, Journal of Machine Learning Research 12 (2011) 2825–2830.
[17] N. Da, The computational case against computational literary studies, Critical Inquiry 45 (2019) 601–639. doi:10.1086/702594.
[18] M. Skeppstedt, M. Ahltorp, K. Kucher, M. Lindström, From word clouds to Word Rain: Revisiting the classic word cloud to visualize climate change texts, Information Visualization (2024). doi:10.1177/14738716241236188.
[19] S. Bird, NLTK: The natural language toolkit, in: Proceedings of the ACL Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics, Association for Computational Linguistics, Stroudsburg, PA, USA, 2002.
[20] V. Danilova, Y. Söderfeldt, Classifying textual genre in historical magazines (1875-1990), in: Proceedings of the 9th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature (LaTeCH-CLfL 2025), Association for Computational Linguistics, Albuquerque, New Mexico, 2025, pp. 160–171. doi:10.18653/v1/2025.latechclfl-1.15.

A. Appendix

As an appendix, we have included the pdf version of the Topic Timelines visualization, as well as screenshots of the ActDisease text database interface.

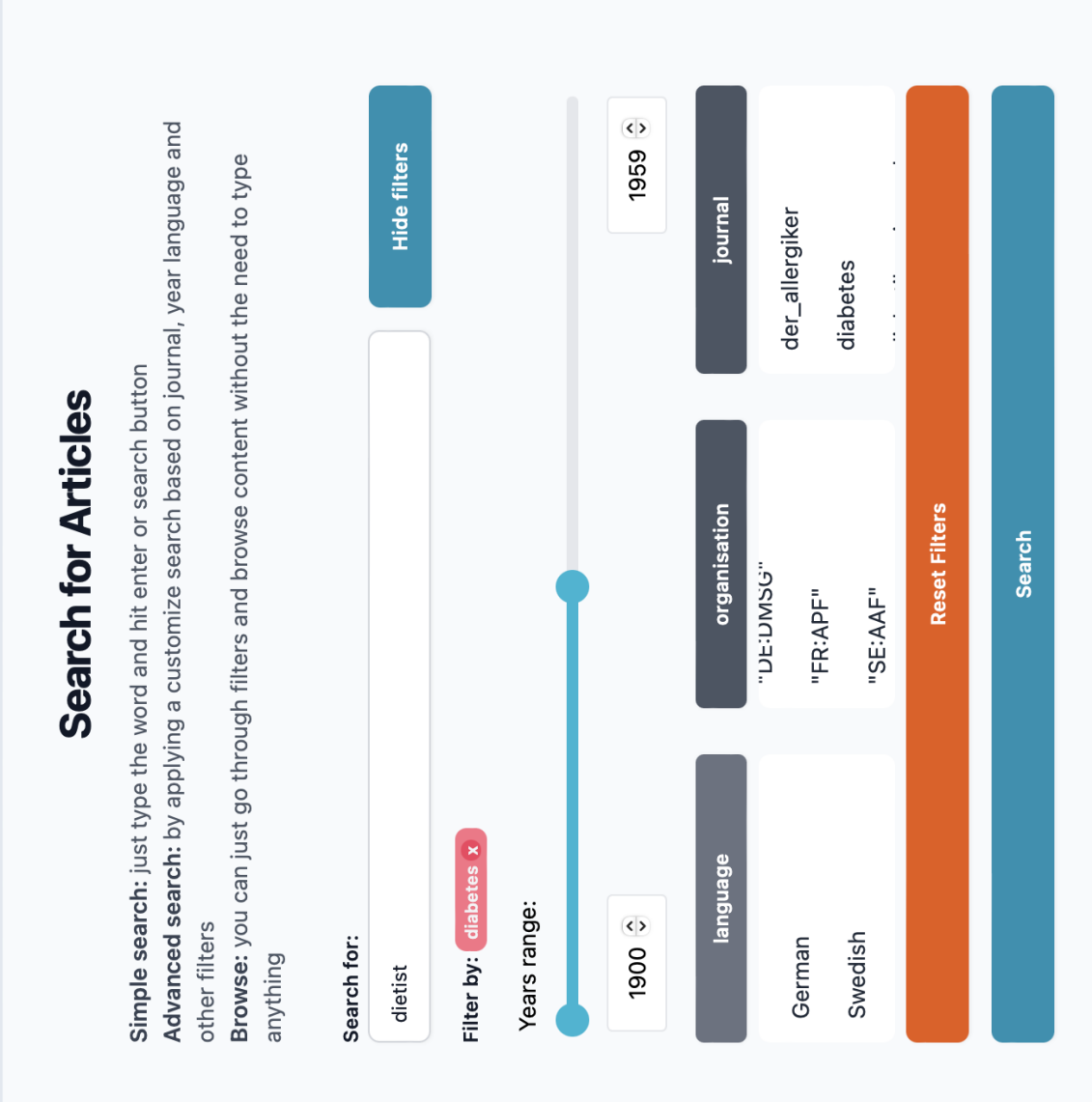


Figure 1: The graphical user interface for searching and filtering the ActDisease database.

Figure 2: The ActDisease database interface showing the results of searching and filtering.

Figure 3: Timeline for the Swedish Diabetes Association member publications.

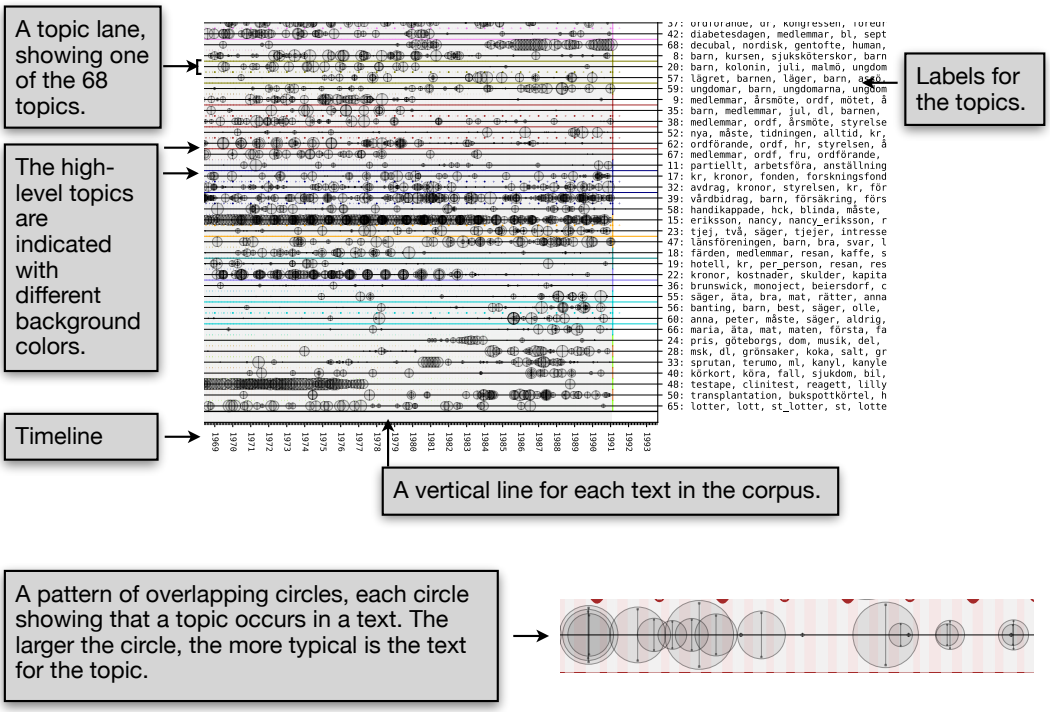


Figure 4: Components of a Topic Timelines graph.

Figure 5: Example of an advertisement, as shown in the user interface for the ActDisease database.

Towards Shared Standards for Pseudonymization of Research Data

Elena Volodina[1,*], Simon Dobnik[2], Therese Lindström Tiedemann[3], Ricardo Muñoz Sánchez[1], Maria Irena Szawerna[1], Lisa Södergård[3] and Xuan-Son Vu[4]

[1]*Språkbanken Text, SFS, University of Gothenburg, Sweden*
[2]*FLOV, University of Gothenburg, Sweden*
[3]*Department of Finnish, Finno-Ugrian and Scandinavian Studies, University of Helsinki, Finland*
[4]*Lund University and DeepTensor AB, Sweden*

Abstract

Pseudonymization has attracted a lot of attention recently due to legislation (e.g. the GDPR), the European Guidelines on Pseudonymization, the increased need for high-quality ethical data for the training of large language models as well as the desire to be able to share data with other researchers. This article introduces key concepts in pseudonymization, summarizes the half-way findings in the intradisciplinary research environment Mormor Karl, and proposes ways to unify and standardize the field of pseudonymization.

Keywords

Open data, GDPR, ethical AI, Mormor Karl, pseudonymization, anonymization, linguistics, Swedish, large language models, privacy

1. Introduction

There is a legal obligation to protect the privacy of data subjects as well as other people mentioned in research data [1, 2]. There exist various techniques to do so, such as encryption, authorization, data minimization, anonymization, pseudonymization [3, 4, 5]. Although none of these approaches can guarantee *absolute protection* of personal privacy [6, 7], they lower the risk of reidentification [8, 7], leading to continuous development of such approaches.

The field of pseudonymization has attracted a lot of attention lately due to legislation like the GDPR [1] and the European Guidelines on Pseudonymization[1], as well as the increased need of high-quality expert data for ethical training of language models [9, 10, 11]. According to the GDPR [1, Art.4:5], pseudonymization is a technique where *Personally Identifiable Information* (PII) have been replaced with substitutes and it is only possible to re-identify a person through additional information, such as name-id keys, which are kept separate from the data. This is a critical difference from the *anonymous* data, where no such keys exist and no reidentification is feasible. This potentially means that once the project destroys the keys, the data should no longer be under the jurisdiction of the GDPR and can be made open to the public. However, the national legal landscape may obstruct that step, in our particular case, the Swedish Ethical Review Authority requires the original data (including the keys) to be preserved for the first ten years after data release; and the Swedish Archives Act, Arkivlagen (SFS 1990:782), protects the documentation (including the above-mentioned keys) from being destroyed in an unauthorized way.

Despite much attention the field is still not unified, and it is challenging to compare the results achieved by different research groups. As a community we need to understand the extent of pseudonym effects on research conclusions, define the tolerance levels and to find a compromise that may be acceptable for both sides - the research in a discipline, and the privacy protection. In the research environment

Huminfra Conference 2025, Stockholm, 12–13 November 2025.
*Corresponding author.
mormor.karl@svenska.gu.se (E. Volodina)

[1]https://www.edpb.europa.eu/system/files/2025-01/edpb_guidelines_202501_pseudonymisation_en.pdf

0. orig.	Hi my name is Deniz Kaya, I live in Sweden in Jämtland and I speak Turkish
1. detect.	Hi my name is Deniz Kaya, I live in in Sweden Jämtland and I speak Turkish
2. label.	Hi my name is @firstname_male.1 @surmale.2, I live in @country.3 in @region.4 and I speak @lang.5
3. pseudo	Hi my name is Alex Bax, I live in Spain in Andalusia and I speak Spanish

Table 1
Example of a sentence containing Personally Identifiable Information (PII) and the processing steps.

group *Mormor Karl* (Eng. 'Grandma Karl')[2] we work on this particular approach to privacy protection, *pseudonymization*, and with a focus on text-based (Swedish) linguistic research data. In the project, we delve into both the methodological and practical issues related to pseudonymization as a way to secure open access to research data [12]. The *practical issues* cover approaches to detect, label and replace personal information - both manually and automatically [e.g. 13]. The *methodological issues* cover, among others, the effects of pseudonymization on research conclusions [e.g. 14, 15, 16]; effectiveness of privacy protection through pseudonymization; as well as the semantic and cultural value of original tokens versus their pseudonyms [e.g. 17, 18, 19]. In this paper, we propose a few directions towards the unification of the field, after shortly outlining the research context for pseudonymization.

2. Pseudonymization: the basics

We define *pseudonymization* as "the process of replacing an individual's personal data with a pseudonym, which is not related to the original data" [12]. An important notion for pseudonymization is *Personally Identifiable Information* (PII). PII is any data that can be used to distinguish, trace or identify an individual, directly, indirectly or in combination with other information sources. PII is conventionally split into categories, such as names, institutions, geographical names and similar. Table 1:2 exemplifies some of PII categories, e.g. @surname, @country, etc. Another related notion is *sensitive information*, such as sexual orientation, political views, religion, ethnical background and medical condition.

Pseudonymization generally includes techniques conventionally divided into several steps, as shown in Table 1. The initial steps comprise detection (Tab.1:1) and labeling (Tab.1:2) of personal (and sensitive) information in unstructured texts, such as essays [20, 21, 22], medical records [23, 24, 25], or court cases [26, 27, 28]. These are followed by the replacement of personal and sensitive information with (neutral) pseudonyms, epithets or codes (Tab.1:3). This process or its parts can be attempted manually [e.g. 20] or automatically [e.g. 29]. It is important to note that, while some studies of automatic pseudonymization treat detection, labelling and replacement separately [13], other approaches merge them together [cf. 30].

2.1. Pseudonymization and linguistics

The early principle in relation to pseudonymization in linguistics has been that we should choose pseudonyms which match *all* linguistic properties of the original, including the number of syllables and other length measures [31]. Recent developments in automatic pseudonymization have often instead stressed both the impossibility of retaining all linguistic properties (e.g. spelling errors and inflectional characteristics) and the need for the pseudonym (cf. the definition above) to be disconnected from the original.

Real names of people (*orthonyms*) are closely related to a person's identity [32], with associations to age, gender, ethnicity, and social background [cf. 33, 34, 35]. While pseudonyms protect the identity of the person, they may have different connotations, thus affecting how the participant is perceived [36] and possibly also how the whole text is interpreted. Wang et al. [37] stress that careful consideration is needed when renaming participants, in order to represent them in an appropriate way, e.g. using Anglo-sounding names for participants with diverse backgrounds implies that the participants' identities and background

[2] https://mormor-karl.github.io/

are not respected. Similarly, placenames (*toponyms*) can be connected to specific historical, cultural and topographical associations [36, 38]. Changing placenames may also affect how the text is perceived. In addition, the language of the placename might be important since in bilingual societies there might be toponyms for one place in both languages and which one you should use could depend on the language you are speaking, hence if you break that norm that might say something about your linguistic proficiency. For instance, in Finland many places have both Finnish and Swedish names, e.g. the capital of Finland is *Helsinki* in Finnish, but *Helsingfors* in Swedish. If you use the Finnish placename when speaking/writing in Swedish this will stand out and a reader/listener might interpret it as (1) not knowing Swedish very well, and/or (2) having certain attitudes regarding dual language placename policy.

But pseudonymization is not only about names, it also covers changing other PIIs in the text. This can involve changing lexical items related to relatives, occupations, health issues, etc. It can also involve changing pronouns and numerals in relation to age, buslines, house numbers etc. Basically, any linguistic item might be affected. This in turn means that the semantic (and pragmatic) context in the text can be affected immensely – *My cousin Tom is 24 years old* is not the same as *My grandma Karl is 27 years old* as we have illustrated in the name of our project.

From the linguistic point of view, we must, therefore, consider how we can retain the original meaning of the text considering lexical semantics but also coreference, contextual semantics and pragmatics. To do this our project uses methods related to linguistic theories from e.g. semantics, pragmatics and grammar. Our results are particularly important for linguistic research, but changes such as PII replacements have profound effects also on other disciplines working on textual data. An analysis in education and social sciences where the connection to certain social groups or regions has been distorted in the data is bound to affect the results [cf. 38].

2.2. Pseudonymization and Natural Language Processing

In automatic pseudonymization (see steps in Table 1) rule-based methods have often been employed [e.g. 39, 23] for the automatic detection and labelling of personal and private information, and this still remains a valid alternative for low-resource settings [40, 41]. However, approaches based on machine learning have been shown to provide the highest performance [42] in some settings. Szawerna et al. [43] argue that the concept of personal or sensitive information and the types of entities that can appear differ between domains, so it is important to be aware that the best approach for a given task or dataset might not necessarily be the same for others. Heterogeneity of classes can also play a major role: a *miscellaneous*-type category (miscellaneous) which encompasses all personal information not covered by other categories is notoriously difficult to detect automatically [44, 13], not to mention issues which then follow as part of replacement.

The most challenging and underexplored step in the automatic pseudonymization process is *pseudonym generation* [45]. This step goes beyond replacing entities with placeholders like `@name`. These pseudonyms should match the context grammatically and semantically to avoid sentences that are non-sensical in context. Besides, in linguistic data it is important to keep as much of the linguistic information as possible, and deciding on what is enough and what is too risky to the person can be extremely difficult, not to mention the possible effects on the usefulness of the data for linguistic research (cf. Section 2.1, *My cousin Tom* vs *My grandma Karl*).

Different approaches have been tried for replacement. These include manual pseudonymization [20], rule-based approaches [46, 23, 29, 21] based on ontologies and entity mapping [47], hierarchical word representations [48], statistical models [49, 30], and machine learning, including Large Language Models (LLMs) [30, 50]. Each of these has its pros and cons. *Manual approaches* are more reliable as far as keeping the semantic integrity of the text is concerned, but they are very time-consuming and risk being inconsistent. Among the automatic approaches, *rule-based approaches* and *statistical models* cannot take into account the semantic context and the common knowledge aspects of the surrounding text, whereas approaches based on *machine learning* and *LLMs* in particular tend to be better where surface semantics are concerned. However, there are some major issues which arise with the use of generative language models. They might rewrite the input text to a more "fluent" version and miss the deeper semantics. This

is problematic as changes must be minimal when it comes to pseudonimyzing linguistic data collected to study language use. Additionally, larger models are often run on external servers, particularly in academic settings. This risks running into legal issues in case the texts cannot be shared due to the nature of the personal information contained within.

3. Unifying strategies: a proposal

3.1. Universal pseudo-tagset

If researchers in different research domains and languages could agree to use the same standard for pseudonymization, similar to the Universal Dependencies initiative [51], it could ensure comparability of datasets, results and it could promote the development of multilingual solutions.

We are currently exploring two approaches: the first one deals with the proposal of a *detailed* pseudo-tagset, that could cover the needs of all research domains [e.g. 43]. The second is the opposite of the first and deals with *reducing all tags to one "personal" tag* [e.g. 52, 13]. Hypothetically, the 'one-tag approach' is feasible for automatic detection of personal information, and having fewer categories to choose between benefits machine-learning approaches. In turn, the 'detailed-tag approach' can be crucial for the selection of appropriate pseudonyms e.g. in rule-based approaches.

Some of the major conceptual and practical challenges when it comes to proposing a universal multi-tag tagset are taxonomy choice and interoperability. Szawerna et al. [43] have shown that existing tagsets vary in terms of types of personal and sensitive information that they cover depending on the genre and domain. A proposed universal tagset would have to account for all of the possibilities and feature a way to expand it in case a new kind of category becomes ubiquitous. Simultaneously, a number of annotated corpora already exist (even if they are not all publically available). Being able to easily map between at least some of the categories would help facilitate the re-annotation to a common standard.

3.2. Testing the effects on research data

Research communities and disciplines need to analyze the potential effects of pseudonymization on their research data as well as on conclusions within their disciplines. This is an important but rather neglected aspect of pseudonymization. Within our project we have done some experiments on effects of pseudonyms on language proficiency assessment. The first one showed no effects on automatic assessment when a first name within a learner essay is changed to a name from another sociocultural background [53]. Testing the same experimental setup with human assessors also showed no clear indication of a correlation between the assessment and the sociocultural associations of the first names that were used [54, 17]. The experiment is currently being extended to include more categories.

Another experiment looked into the linguistic analysis of PII strings to uncover what information may be lost if the original strings were replaced with pseudonyms [19, 55, 18]. Our analysis shows that a non-negligible number of PII are misspelled - information that is lost in the automatic pseudonymization process. However, misspellings carry important information, e.g. they can be linked to the native language of a learner e.g. *Danska* which literally means 'Danish' but where the intention is clearly Denmark and it is influenced by the Finnish name *Tanska* 'Denmark'. Pseudonymization will lose the connection to the mother tongue and also the knowledge implied in the use of <d> instead of <t> in the spelling.

We suggest to standardize the practice of testing effects of pseudonymization on research conclusions through a few typical tasks within the target domain and report the results.

3.3. Testing effectiveness of pseudonymization

Advances in machine learning (ML) privacy and security reveal critical vulnerabilities in deep neural networks deployed in sensitive domains. There are three interconnected threat landscapes: (1) adversarial attacks that manipulate model behavior through input perturbations, (2) reidentification attacks that extract sensitive training data through model inversion and membership inference, and (3) motivated

intruders - individuals who use publicly available or background information to attempt re-identifying individuals in pseudonymized text.

Collectively, they demonstrate that modern ML systems face severe privacy and security risks, with re-identification succeeding against 50–90% of vulnerable models [56], and anonymized datasets showing 10–40% re-identification rates under motivated intruder scenarios [57]. The convergence of these threats necessitates integrated defense strategies combining robust architectures, privacy-enhancing computations, and rigorous validation protocols.

Pseudonymization techniques show limited resilience against motivated intruders, as empirical studies highlight vulnerabilities in datasets thought to be de-identified when public data sources are exploited. This emphasizes the urgent need for advanced pseudonymization methods to mitigate such risks and robust methods to diagnose such vulnerabilities. The increasing focus on privacy-enhancing technologies under frameworks like the GDPR [1] and the EU AI Act [58] reflects the importance of regulatory alignment in strengthening data protection. In connection to that we encourange the community to standardize the use of reidentification tests as a way to evaluate effectiveness of pseudonymization, in addition to the standard preformance tests.

3.4. Customizable pseudonymization

Another issue to consider is whether pseudonymization should be *customizable* to different types of users and use cases. Imagine a *sociolinguist*, who will look for linguistic details of interest to describe a certain variety or issues in relation to language and power, or *a data scientist*, who will not inspect the actual data manually but as properties of a dataset as a whole, or *a forensic linguist* who will try to identify the person behind the writing. If we start applying different types of pseudonymization to different use cases and users, we *need to standardize different pseudonymization criteria and methods* so that they can be objectively reported in research (e.g. we used data X with pseudonymization Y). Thorough analysis of re-identification risks is extremely urgent in this case since having access to different pseudonymized versions of the same data increases the risk of re-identification.

A good way forward is to explore a dynamic approach for pseudonymization that would identify per text and context whether the text can be matched to a situation within the context. Hence, to achieve this one can simply compare the text with the context: is the psedonymized text consistent with the context of a research domain it is used for? The task could perhaps be treated as *Natural Language Inference Task (NLI)* [59].

Another implication of this approach is that pseudonymization would then be a tool to create several versions of a given dataset which would be adjusted to a given task. However, the approach still does not answer the question of possible re-identification in cases where several pseudonymized versions of the same text could be merged together to reconstruct the information from the original text.

3.5. Evaluation benchmarks

Another field-unifying strategy that we look into relates to the possibility of organizing shared tasks on the topic of automatic detection and pseudonymization of personal information. The community needs standardized evaluation benchmarks for pseudonymization tasks, and we expect the data from shared tasks to fulfil this function. The immediate concerns are:

(a) What data to use – much of the original data containing authentic personal information is under protection and cannot be released or used for model training. The existing publically available PII-annotated corpora [26, 60, 61, 25, 62, 63] are likely to have already been a part of the training data for LLMs, as well as cover only limited number of languages. *Synthetic data* [e.g. 64] may be good enough for development of automated detection methods, as has been indicated by Vakili et al. [65], but, since it is not authentic human-produced data it should not be used for anything beyond training models.

(b) The second concern is how to perform *automatic evaluation of the pseudonymization step*, where there are no agreed-upon standard metrics. A part of the ongoing work in the project is aimed at testing

the validity of various automatic evaluation approaches, attempting to approximate the human judgements of grammatical and semantic acceptability [cf. 66].

Our approach to circumventing the legal limitations and ethical concerns surrounding the use of authentic personal information for a shared task consist of creating a corpus of fictive texts. Unlike purely synthetic data, our texts are written by human respondents. However, they are not written about natural persons, but about fictive characters invented for the sake of writing. This minimizes the risk to any natural person and approximates the way that authentic texts are written. Our data collection so far consists of fictive personal stories and fictive legal case descriptions, but more domains (e.g. medical, social media) are planned.

3.6. Standardized venues

There is a clear need for venues for meetings both within disciplines and in interdisciplinary groups to discuss pseudonymization challenges and share findings. We have initiated two venues for meeting researchers working with similar questions within *computational linguistics, computer science and privacy*, in particular: (a) the CALD-pseudo workshop[3] with its first edition at EACL 2024 and which we hope to make a standard recurrent venue on this topic. (b) the AI Trust workshop[4] with the first edition in 2024 at the WASP conference in Gothenburg. This workshop has a slightly broader focus than CALD-pseudo, but both workshops indicated signficant interest and an expressed need to meet and discuss these issues both within and across disciplines. In 2025 we also reached out to *linguists* working on the Swedish language and held a workshop at Svenskans beskrivning 40 (2025)[5] which similarly proved a joint sense of a need for more research on pseudonymization techniques and their effects on linguistic research.

Our experience with interacting with different audiences through these workshops shows that there is an increasing need for unified procedures and guidelines related to dealing with pseudonymization of research data that does not only affect linguistics and computational linguistics but also has implications for other domains, some of which may not strictly use typical natural language processing data.

4. Concluding remarks and future outlook

There are still many open issues in relation to pseudonymization of research data and what it means for our disciplines as well as for research participants (e.g. writers of essays, people who are interviewed) or people that are mentioned in our research data. The results of our research have clear benefits to research infrastructures on a practical level, and important implications for research on the methodological level.

Research on pseudonym generation is in its initial stages. A promising solution for the semantics-based issues might be to *generate fake or non-existent entities*. That is, names, cities, etc. that would look like real ones but not contain any semantic or pragmatic value for the reader, although preserving the grammatical features. However, both real and fake names with similar structure can have very different associations and reference.

Research on effects of pseudonymization on research conclusions has hardly begun and there is still much to be explored in relation to automatization, bias and privacy preservation. Further analysis is necessary, especially in relation to actual research questions on actual research data for us to ascertain that our research will still be reliable after pseudonymization, that the rights of our participants will be protected both in relation to privacy preservation and in terms of their right to their data, their culture and the correctness of research findings.

One open question is how to handle 'privacy guarantees' when research data comes in *several modalities*, i.e. not only as text datasets, but also speech/audio, video, pictures. In our project we are focusing on

[3] https://mormor-karl.github.io/events/CALD-pseudo/#cald-pseudo-workshop-at-eacl-2024
[4] https://mormor-karl.github.io/events/AITrust-Workshop/
[5] https://www.su.se/institutionen-for-svenska-och-flersprakighet/forskning/konferenser-och-seminarier/konferens-2025-svenskans-beskrivning-40-1.730200

written (and transcribed) data, but there is a definite need to look at other modalities also and this too will need to be done in relation to different research disciplines.

Limitations

The work is focused on text modality of research data only, which means we do not look into other types of privacy protection, where video, audio, graphics (including, for example, handwritten versions of texts in our collection) present further challenges.

Ethics Statement

To conduct this research, we follow all legal and ethical practices and are in constant contact with university lawyers.

Acknowledgements

This work has been funded through the research environment grant from the Swedish Research Council: *Grandma Karl is 27 years old: Automatic pseudonymization of research data* (nr.2022-02311).

References

[1] E. EU Commission, General data protection regulation., Official Journal of the European Union, 59, 1-88., 2016. URL: https://gdpr-info.eu/.

[2] M. R. C. MRC, GDPR Guidance note 5: Identifiability, anonymisation and pseudonymisation, 2019. URL: https://mrc.ukri.org/documents/pdf/gdpr-guidance-note-5-identifiability-anonymisation-and-pseudonymisation/, (Accessed 2025-09-26).

[3] ENISA, Privacy Enhancing Technologies: Evolution and State of the Art. A Community Approach to PETs Maturity Assessment, 2017.

[4] ENISA, A tool on Privacy Enhancing Technologies (PETs) knowledge management and maturity assessment, 2018. URL: https://www.enisa.europa.eu/publications/pets-maturity-tool, (Accessed 2025-09-26).

[5] G. Danezis, J. Domingo-Ferrer, M. Hansen, J.-H. Hoepman, D. Le Métayer, R. Tirtea, S. Schiffner, Privacy and Data Protection by Design – from policy to engineering, 2014. URL: https://www.enisa.europa.eu/publications/privacy-and-data-protection-by-design.

[6] L. Rocher, J. M. Hendrickx, Y.-A. De Montjoye, Estimating the success of re-identifications in incomplete datasets using generative models, Nature communications 10 (2019) 1–9.

[7] L. G. G. Charpentier, P. Lison, Re-identification of de-identified documents with autoregressive infilling, in: W. Che, J. Nabende, E. Shutova, M. T. Pilehvar (Eds.), Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Vienna, Austria, 2025, pp. 1192–1209. URL: https://aclanthology.org/2025.acl-long.60/.

[8] B. Manzanares-Salor, D. Sanchez, P. Lison, Evaluating the disclosure risk of anonymized documents via a machine learning-based re-identification attack, Data Mining and Knowledge Discovery 38 (2024) 4040–4075.

[9] Z. Ji, Y. Shen, K. R. Koedginer, J. Lin, Enhancing the de-identification of Personally Identifiable Information in educational data, Journal of Educational Data Mining 17(2) (2025) 55–85. URL: https://doi.org/10.5281/zenodo.17114271.

[10] I. Pilán, B. Manzanares-Salor, D. Sánchez, P. Lison, Truthful text sanitization guided by inference attacks, arXiv preprint arXiv:2412.12928 (2025). URL: https://doi.org/10.48550/arXiv.2412.12928.

[11] J. Zhang, Z. Tian, M. Zhu, Y. Song, T. Sheng, S. Yang, Q. Du, X. Liu, M. Huang, D. Li, DYNTEXT: semantic-aware dynamic text sanitization for privacy-preserving LLM inference, in: Findings of the Association for Computational Linguistics: ACL 2025, 2025, pp. 20243–20255.

[12] E. Volodina, S. Dobnik, T. Lindström Tiedemann, V. Xuan-Son, Grandma Karl is 27 years old - research agenda for pseudonymization of research data, in: Proceedings of 2023 IEEE Ninth International Conference on Big Data Computing Service and Applications (BigDataService), the 2023 Workshop on Big Data and Machine Learning with Privacy Enhancing Tech, 2023.

[13] M. I. Szawerna, S. Dobnik, R. Muñoz Sánchez, E. Volodina, The devil's in the details: the detailedness of classes influences personal information detection and labeling, in: R. Johansson, S. Stymne (Eds.), Proceedings of the Joint 25th Nordic Conference on Computational Linguistics and 11th Baltic Conference on Human Language Technologies (NoDaLiDa/Baltic-HLT 2025), University of Tartu Library, Tallinn, Estonia, 2025, pp. 697–708. URL: https://aclanthology.org/2025.nodalida-1.70/.

[14] L. Södergard, Deltagare 1, K1989, Lova eller Latife – hur forskare benämner personer som förekommer i forskningsmaterialet, in: Svenskan i Finland, submitted.

[15] L. Södergard, Pseudonymisering av orter, skolor och organisationer. Hur gör språkforskare i praktiken?, in: Svenskans beskrivning 40, Stockholms universitet, in progress.

[16] R. Muñoz Sánchez, S. Dobnik, M. I. Szawerna, T. Lindström Tiedemann, E. Volodina, Did the names I used within my essay affect my score? Diagnosing name biases in automated essay scoring, in: E. Volodina, D. Alfter, S. Dobnik, T. Lindström Tiedemann, R. Muñoz Sánchez, M. I. Szawerna, X.-S. Vu (Eds.), Proceedings of the workshop on Computational Approaches to Language Data Pseudonymization (CALD-pseudo 2024), 2024, pp. 81–91.

[17] T. Lindström Tiedemann, L. Södergard, R. Muñoz Sánchez, S. Dobnik, M. I. Szawerna, Names, pseudonyms and biases in language assessment, TBA (in progress).

[18] T. Lindström Tiedemann, L. Södergard, E. Volodina, S. Dobnik, M. Szawerna, R. Muñoz Sánchez, X.-S. Vu, Om mormor Karl sägs vara 27 år gammal, vad säger det om skribenten? En presentation om att identifiera och ersätta identifierande element i språkvetenskapliga forskningsdata [=If Grandma Karl is said to be 27 years old, what does that say about the writer? A presentation about identifying and replacing identying elements in linguistic research data], in: Svenskans beskrivning 40, Stockholms universitet, in progress.

[19] L. Södergard, T. Lindström Tiedemann, Att ansvarsfullt skydda och dölja identitet i språkforskning [=to protect and obscure identity responsibly in linguistic research], in: Kielitieteen paivat Spraakvetenskapsdagarna The Finnish Conference of Linguistics, Helsinki 12–14 May 2025, 2025.

[20] B. Megyesi, L. Granstedt, S. Johansson, J. Prentice, D. Rosén, C.-J. Schenström, G. Sundberg, M. Wirén, E. Volodina, Learner Corpus Anonymization in the Age of GDPR: Insights from the Creation of a Learner Corpus of Swedish, in: Proceedings of the 7th NLP4CALL, Swedish Language Technology Conference, SLTC 2018, 2018, pp. 47–56.

[21] E. Volodina, Y. A. Mohammed, S. Derbring, A. Matsson, B. Megyesi, Towards privacy by design in learner corpora research: A case of on-the-fly pseudonymization of Swedish learner essays, in: Proceedings of the 28th International Conference on Computational Linguistics, 2020.

[22] M. I. Szawerna, S. Dobnik, R. Muñoz Sánchez, T. Lindström Tiedemann, E. Volodina, Detecting personal identifiable information in Swedish learner essays, in: E. Volodina, D. Alfter, S. Dobnik, T. Lindström Tiedemann, R. Muñoz Sánchez, M. I. Szawerna, X.-S. Vu (Eds.), Proceedings of the Workshop on Computational Approaches to Language Data Pseudonymization (CALD-pseudo 2024), Association for Computational Linguistics, St. Julian's, Malta, 2024, pp. 54–63. URL: https://aclanthology.org/2024.caldpseudo-1.7/.

[23] H. Dalianis, Pseudonymisation of Swedish electronic patient records using a rule-based approach, in: L. Ahrenberg, B. Megyesi (Eds.), Proceedings of the Workshop on NLP and Pseudonymisation, Linköping Electronic Press, Turku, Finland, 2019, pp. 16–23. URL: https://aclanthology.org/W19-6503/.

[24] P. Ngo, M. Tejedor, T. Olsen Svenning, T. Chomutare, A. Budrionis, H. Dalianis, Deidentifying

a Norwegian clinical corpus - an effort to create a privacy-preserving Norwegian large clinical language model, in: E. Volodina, D. Alfter, S. Dobnik, T. Lindström Tiedemann, R. Muñoz Sánchez, M. I. Szawerna, X.-S. Vu (Eds.), Proceedings of the Workshop on Computational Approaches to Language Data Pseudonymization (CALD-pseudo 2024), Association for Computational Linguistics, St. Julian's, Malta, 2024, pp. 37–43. URL: https://aclanthology.org/2024.caldpseudo-1.5/.

[25] M. Marimon, A. Gonzalez-Agirre, A. Intxaurrondo, J. A. L. Martin, M. Villegas, Automatic De-Identification of Medical Texts in Spanish: the MEDDOCAN Track, Corpus, Guidelines, Methods and Evaluation of Results (2019).

[26] I. Pilán, P. Lison, L. Øvrelid, A. Papadopoulou, D. Sánchez, M. Batet, The text anonymization benchmark (TAB): A dedicated corpus and evaluation framework for text anonymization, Computational Linguistics 48 (2022) 1053–1101.

[27] M. Sierro, B. Altuna, I. Gonzalez-Dios, Automatic detection and labelling of personal data in case reports from the ECHR in Spanish: Evaluation of two different annotation approaches, in: E. Volodina, D. Alfter, S. Dobnik, T. Lindström Tiedemann, R. Muñoz Sánchez, M. I. Szawerna, X.-S. Vu (Eds.), Proceedings of the Workshop on Computational Approaches to Language Data Pseudonymization (CALD-pseudo 2024), Association for Computational Linguistics, St. Julian's, Malta, 2024, pp. 18–24. URL: https://aclanthology.org/2024.caldpseudo-1.3/.

[28] T. Allard, L. Béziaud, S. Gambs, Publication of court records: circumventing the privacy-transparency trade-off, in: AICOL 2020 - 11th International Workshop on Artificial Intelligence and the Complexity of Legal Systems, in conjunction with JURIX 2020, Virtual, Czech Republic, 2020. URL: https://inria.hal.science/hal-03225201, a version of this work was presented at the Law and Machine Learning workshop at ICML 2020 (no proceeding).

[29] E. Eder, U. Krieg-Holz, U. Hahn, De-identification of emails: Pseudonymizing privacy-sensitive data in a German email corpus, in: R. Mitkov, G. Angelova (Eds.), Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019), INCOMA Ltd., Varna, Bulgaria, 2019, pp. 259–269. URL: https://aclanthology.org/R19-1030/.

[30] O. Yermilov, V. Raheja, A. Chernodub, Privacy- and utility-preserving NLP with anonymized data: A case study of pseudonymization, in: A. Ovalle, K.-W. Chang, N. Mehrabi, Y. Pruksachatkun, A. Galystan, J. Dhamala, A. Verma, T. Cao, A. Kumar, R. Gupta (Eds.), Proceedings of the 3rd Workshop on Trustworthy Natural Language Processing (TrustNLP 2023), Association for Computational Linguistics, Toronto, Canada, 2023, pp. 232–241. URL: https://aclanthology.org/2023.trustnlp-1.20/.

[31] E. Callegari, A. Sólmundsdóttir, A. K. Ingason, Preserving Privacy in Small Communities: Tailored Anonymization Techniques for Icelandic Conversational Data, in: CLARIN Annual Conference Proceedings, 2024, p. 121.

[32] T. Ainiala, J.-O. Östman, Introduction: Socio-onomastics and pragmatics, in: Socio-onomastics, John Benjamins Publishing Company, 2017, pp. 1–18.

[33] E. Aldrin, Names as resources for gendering: Trends within the field, Nordic Journal of Socio-Onomastics 5 (2025) 5–32.

[34] E. Aldrin, Vad säger väl ett namn?: Reflektioner kring teorin om markerade namn utifrån exemplet etniska konnotationer till förnamn, in: Norna-rapporter, volume 100, NORNA-förlaget, 2023, pp. 57–79.

[35] M. Frändén, " vi bestämde oss för att skriva namnet på ett svenskt sätt": Förnamnsval i sverigefinska familjer, Studia Anthroponymica Scandinavica (2015) 75–138.

[36] J. Heaton, "* pseudonyms are used throughout": A footnote, unpacked, Qualitative Inquiry 28 (2022) 123–132.

[37] S. Wang, J. M. Ramdani, S. Sun, P. Bose, X. Gao, Naming research participants in qualitative language learning research: Numbers, pseudonyms, or real names?, Journal of language, identity & education (2024) 1–14.

[38] J. L. Seelig, Place anonymization as rural erasure? A methodological inquiry for rural qualitative scholars, International Journal of Qualitative Studies in Education 34 (2021) 857–870.

[39] P. Accorsi, N. Patel, C. Lopez, R. Panckhurst, M. Roche, Seek&Hide: Anonymising a French SMS corpus using natural language processing techniques, Linguisticae Investigationes 35 (2012) 163–180. doi:10.1075/li.35.2.03acc.

[40] R. Blokland, N. Partanen, M. Rießler, A pseudonymisation method for language documentation corpora: An experiment with spoken Komi, in: T. A. Pirinen, F. M. Tyers, M. Rießler (Eds.), Proceedings of the Sixth International Workshop on Computational Linguistics of Uralic Languages, Association for Computational Linguistics, Wien, Austria, 2020, pp. 1–8. URL: https://aclanthology.org/2020.iwclul-1.1/.

[41] N. Ilinykh, M. I. Szawerna, "I need more context and an English translation": Analysing how LLMs identify personal information in Komi, Polish, and English, in: Š. A. Holdt, N. Ilinykh, B. Scalvini, M. Bruton, I. N. Debess, C. M. Tudor (Eds.), Proceedings of the Third Workshop on Resources and Representations for Under-Resourced Languages and Domains (RESOURCEFUL-2025), University of Tartu Library, Estonia, Tallinn, Estonia, 2025, pp. 165–178. URL: https://aclanthology.org/2025.resourceful-1.32/.

[42] V. Yogarajan, B. Pfahringer, M. Mayo, A review of automatic end-to-end de-identification: Is high accuracy the only metric?, Applied Artificial Intelligence 34 (2020) 251–269. URL: https://doi.org/10.1080/08839514.2020.1718343.

[43] M. I. Szawerna, S. Dobnik, T. Lindström Tiedemann, R. M. Sánchez, X.-S. Vu, E. Volodina, Pseudonymization categories across domain boundaries, in: Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), 2024, pp. 13303–13314.

[44] A. Papadopoulou, Y. Yu, P. Lison, L. Øvrelid, Neural text sanitization with explicit measures of privacy risk, in: Y. He, H. Ji, S. Li, Y. Liu, C.-H. Chang (Eds.), Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), Association for Computational Linguistics, Online only, 2022, pp. 217–229. URL: https://aclanthology.org/2022.aacl-main.18/.

[45] P. Lison, I. Pilán, D. Sanchez, M. Batet, L. Øvrelid, Anonymisation models for text data: State of the art, challenges and future directions, in: C. Zong, F. Xia, W. Li, R. Navigli (Eds.), Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), Association for Computational Linguistics, Online, 2021, pp. 4188–4203. URL: https://aclanthology.org/2021.acl-long.323/.

[46] A. Alfalahi, S. Brissman, H. Dalianis, Pseudonymisation of Personal Names and other PHIs in an Annotated Clinical Swedish Corpus, in: Proceedings of the Third Workshop on Building and Evaluating Resources for Biomedical Text Mining (BioTxtM 2012) held in conjunction with LREC 2012, 2012. URL: https://api.semanticscholar.org/CorpusID:6387546.

[47] A. W. Olstad, A. Papadopoulou, P. Lison, Generation of replacement options in text sanitization, in: T. Alumäe, M. Fishel (Eds.), Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa), University of Tartu Library, Tórshavn, Faroe Islands, 2023, pp. 292–300. URL: https://aclanthology.org/2023.nodalida-1.30/.

[48] O. Feyisetan, T. Diethe, T. Drake, Leveraging Hierarchical Representations for Preserving Privacy and Utility in Text , in: 2019 IEEE International Conference on Data Mining (ICDM), IEEE Computer Society, Los Alamitos, CA, USA, 2019, pp. 210–219. doi:10.1109/ICDM.2019.00031.

[49] D. Simancek, V. V. Vydiswaran, Handling name errors of a BERT-based de-identification system: Insights from stratified sampling and Markov-based pseudonymization, in: E. Volodina, D. Alfter, S. Dobnik, T. Lindström Tiedemann, R. Muñoz Sánchez, M. I. Szawerna, X.-S. Vu (Eds.), Proceedings of the Workshop on Computational Approaches to Language Data Pseudonymization (CALD-pseudo 2024), Association for Computational Linguistics, St. Julian's, Malta, 2024, pp. 1–7. URL: https://aclanthology.org/2024.caldpseudo-1.1/.

[50] S. Hou, R. Shang, Z. Long, X. Fu, Y. Chen, A general pseudonymization framework for cloud-based llms: Replacing privacy information in controlled text generation, 2025. URL: https://arxiv.org/abs/2502.15233.
[51] M.-C. de Marneffe, C. D. Manning, J. Nivre, D. Zeman, Universal dependencies, Computational Linguistics 47 (2021) 255–308. doi:10.1162/coli_a_00402.
[52] M. I. Szawerna, S. Dobnik, R. M. Sánchez, T. Lindström Tiedemann, E. Volodina, Detecting personal identifiable information in swedish learner essays, in: Proceedings of the Workshop on Computational Approaches to Language Data Pseudonymization (CALD-pseudo 2024), 2024, pp. 54–63.
[53] R. Muñoz Sánchez, S. Dobnik, M. I. Szawerna, T. Lindström Tiedemann, E. Volodina, Did the names i used within my essay affect my score? diagnosing name biases in automated essay scoring, in: Proceedings of the Workshop on Computational Approaches to Language Data Pseudonymization (CALD-pseudo 2024), 2024, pp. 81–91.
[54] R. Muñoz Sánchez, S. Dobnik, M. I. Szawerna, T. Lindström Tiedemann, E. Volodina, Name biases in automated essay assessment, in: International congress of onomastic sciences, ICOS, Helsinki, 19–23 August 2024, 2024.
[55] T. Lindström Tiedemann, L. Södergard, E. Volodina, S. Dobnik, M. Szawerna, R. Munoz Sanchez, X.-S. Vu, En presentation om att ersätta identifierande element i språkvetenskapliga forskningsdata, in: Workshop Pseudonymisering inom språkvetenskap, Svenskans beskrivning 40, Stockholm, 26 May 2025, 2025.
[56] M. Rigaki, S. Garcia, A survey of privacy attacks in machine learning, ACM Computing Surveys 56 (2023) 1–34.
[57] M. Aerni, J. Zhang, F. Tramèr, Evaluations of machine learning privacy defenses are misleading, in: Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security, 2024, pp. 1271–1284.
[58] Regulation 2024/1689, The EU Artificial Intelligence Act (2024/1689), Official Journal of the European Union, L series, 2016. URL: https://eur-lex.europa.eu/eli/reg/2024/1689/oj.
[59] S. Chatzikyriakidis, R. Cooper, S. Dobnik, S. Larsson, An overview of natural language inference data collection: The way forward?, in: C. Gardent, C. Retoré (Eds.), Proceedings of IWCS 2017: 12th International Conference on Computational Semantics, Workshop on Computing Natural Language Inference, Association for Computational Linguistics, Montpellier, France, 2017, pp. 1–6. URL: http://www.aclweb.org/anthology/W/W17/#7200.
[60] A. Stubbs, C. Kotfila, Özlem Uzuner, Automated systems for the de-identification of longitudinal clinical narratives: Overview of 2014 i2b2/uthealth shared task track 1, Journal of Biomedical Informatics 58 (2015) S11–S19. URL: https://doi.org/10.1016/j.jbi.2015.06.007.
[61] A. Stubbs, M. Filannino, Özlem Uzuner, De-identification of psychiatric intake records: Overview of 2016 cegs n-grid shared tasks track 1, Journal of Biomedical Informatics 75 (2017) S4–S18. URL: https://doi.org/10.1016/j.jbi.2017.06.011, supplement: A Natural Language Processing Challenge for Clinical Records: Research Domains Criteria (RDoC) for Psychiatry.
[62] M. Marimon, A. Gonzalez-Agirre, A. Intxaurrondo, H. Rodríguez, J. A. Lopez Martin, M. Villegas, M. Krallinger, MEDDOCAN corpus: gold standard annotations for Medical Document Anonymization on Spanish clinical case reports , 2020. URL: https://doi.org/10.5281/zenodo.4279323.
[63] L. Holmes, Cleaned Repository of Annotated PII, https://www.kaggle.com/datasets/langdonholmes/cleaned-repository-of-annotated-pii, 2024. [Accessed 18-09-2025].
[64] AI4Privacy, ai4privacy/pii-masking-300k · Datasets at Hugging Face — huggingface.co, https://huggingface.co/datasets/ai4privacy/pii-masking-300k, 2024. [Accessed 05-09-2025].
[65] T. Vakili, A. Henriksson, H. Dalianis, End-to-end pseudonymization of fine-tuned clinical bert models: Privacy preservation with maintained data utility, BMC Medical Informatics and Decision Making 24 (2024) 162.
[66] E. Eder, U. Krieg-Holz, U. Hahn, De-Identification of Emails: Pseudonymizing Privacy-Sensitive Data in a German Email Corpus, in: R. Mitkov, G. Angelova (Eds.), Proceedings of the International

Conference on Recent Advances in Natural Language Processing (RANLP 2019), INCOMA Ltd., Varna, Bulgaria, 2019, pp. 259–269. URL: https://aclanthology.org/R19-1030.

DigiCURE: Building a Digital Humanities Infrastructure for Preserving and Studying At-Risk Cultural Heritage

Jonathan Westin[1], Cecilia Lindhé[1], Daniel Brodén[1], Gunnar Almevik[1], Matteo Tomasini[1]

[1] *University of Gothenburg, Box 200, Gothenburg, 405 30, Sweden*

Abstract

The preservation of cultural heritage has become an urgent societal, policy, and scientific priority in the context of climate change, armed conflicts, and rapid urbanisation. Monuments, archaeological sites, and fragile materials are increasingly at risk of loss or irreversible damage. In response, DigiCURE (Digital Cultural Resilience and Protection) establishes a technologically advanced and institutionally anchored research infrastructure dedicated to the digitisation, preservation, and analysis of endangered heritage. The aim of this paper is to present the DigiCURE research infrastructure, which provides high-quality tools, expertise, and training to support sustainable digital preservation through multimodal documentation, spatial visualisation, and data modelling. Its online platform enables researchers, heritage professionals, and the public to explore, analyse, and annotate complex multimodal datasets with AI-assisted methods, even in cases where physical access is no longer possible. Developed within the framework of the Gothenburg Research Infrastructure in Digital Humanities (GRIDH), DigiCURE functions as both a national and international hub for innovative research, ensuring the long-term accessibility and resilience of vulnerable cultural heritage.

Keywords

Digital cultural heritage, research infrastructure, multimodality

Introduction

Cultural heritage is an indispensable resource for many disciplines and plays a vital role in shaping a shared historical consciousness in increasingly fragmented and globalised societies. Yet around the world, cultural resources face growing threats: archaeological sites erode under rising seas and extreme weather; fragile materials deteriorate due to shifting environmental conditions; museums, archives, and historic monuments are damaged, destroyed, or displaced by war; and increased tourism places additional stress on vulnerable sites. These developments underscore the urgent need for reflective, resilient, and accessible preservation strategies to safeguard information for research [1] [2]. Hence, Swedish and European policy frameworks increasingly recognise the importance of digitisation as a crucial component in preserving cultural heritage. In line with EU recommendations, all member states have committed to digitising at-risk monuments and major sites by 2030, as outlined in the 2019 EU declaration on digital cultural heritage and reaffirmed in the Swedish National Heritage Board's digital strategy (2024).

While there is widespread consensus on the importance of digitising cultural heritage and that digital methods offer powerful means of documentation and preservation [3] [4] [5] [6] [7], the research community faces notable gaps as many disciplines and institutions lack the necessary infrastructure, technical expertise, and sustainable workflows [8]. Furthermore, the multimodal and diverse nature of the data required to accurately capture heritage sites or monuments often leads to the datasets being stored in disparate silos. This fragmentation hinders more in-depth and comprehensive analysis as well as the application of computational approaches. Consequently, digitisation efforts frequently result in

Huminfra Conference 2025, Stockholm, 12-13 November 2025.

jonathan.westin@lir.gu.se (J. Westin); cecilia.lindhe@lir.gu.se (C. Lindhé); daniel.broden@lir.gu.se (D. Brodén); gunnar.almevik@conservation.gu.se (G. Almevik); matteo.tomasini@lir.gu.se (M. Tomasini);

digital resources that fall short of researchers' needs concerning quality, depth, and focus [9] [10] [11] [12].

As a response, the Gothenburg Research Infrastructure in Digital Humanities (GRIDH) and the Department of Conservation (DC) at the University of Gothenburg established an infrastructure to address these critical gaps while advancing digitally supported research across the humanities and a range of related fields. The purpose of this initiative was to provide a long-term, research-oriented infrastructure that supports high-quality digitisation, preservation, reuse, and analysis of at-risk heritage materials.

Multimodal Research Infrastructure

The aim of this paper is to present DigiCURE (Digital Cultural Resilience and Protection), a research infrastructure designed to address these gaps by providing a multimodal platform, advanced digitisation tools, and expertise for the sustainable preservation and analysis of at-risk cultural heritage. DigiCURE offers an institutionally embedded infrastructure built around three core components:

Multimodal platform: Development and maintenance of a robust multimodal platform that facilitates digital preservation and remote access to endangered cultural assets. The platform integrates various data types – including high-resolution images, 3D models, and metadata – to ensure comprehensive documentation and accessibility for research purposes and public engagement.

Multimodal digitisation: Provision of cutting-edge software solutions for digitisation, metadata structuring, and secure archiving. DigiCURE also equips researchers with high-resolution imaging, 3D scanning, and digital reconstruction tools to support detailed documentation and analysis of heritage materials.

Multimodal expertise: Facilitating expert consultation services and providing open-source resources to build capacity in digital preservation practices. DigiCURE ensures adherence to best practices and standards, promoting responsible and sustainable stewardship of cultural heritage. Built from the ground up for long-term sustainability, DigiCURE offers a growing online platform with tools, expertise, and workflows that promote accessibility, interoperability, and analytical depth. It provides a solid technical and methodological foundation for producing interoperable, analysis-ready data, ensuring that at-risk heritage material can be made reliably accessible and reusable for researchers, institutions, and the public.

Today, the DigiCURE infrastructure offers scholarly access to a range of multimodal resources, combining advanced technologies with user-centric design [13]. The infrastructure is hosted and developed by GRIDH, a cross-university research infrastructure at the University of Gothenburg established in 2015 to support digital tools, methods, and platforms in humanities research and to engage in Swedish and European infrastructure consortia. Consequently, DigiCURE's resources and services are published and maintained through GRIDH's resource portal, ensuring long-term visibility, accessibility, and usability. It brings together a team of senior experts in research infrastructure and software engineers with complementary expertise.

Context-Sensitive Design

Integrated Methodology

DigiCURE's methodological design is grounded in the understanding that digitising cultural heritage is not a neutral technical procedure, but a critical and epistemologically charged practice [14] [15]. As Drucker [16] argues, the systems we build to structure and access cultural data actively shape how knowledge is produced, interpreted, and valued. Svensson [17] similarly highlights the importance of "humane infrastructures" that embed interpretive nuance, scholarly judgment, and cultural accountability into the design of digital systems. DigiCURE builds on these insights by approaching heritage as a form of information organisation – an active process of selecting, classifying, and preserving what holds cultural significance for specific communities. This perspective aligns with Harrison's [18] view of heritage as a present-day practice that reflects both contemporary systems of order and the urgency to safeguard identities and their tangible or intangible expressions under threat.

DigiCURE treats digital cultural heritage as a socially embedded, ethically aware, and critically informed practice by involving users as active participants rather than passive recipients.

Hence, by building on established models for interdisciplinary collaboration [19] [20] it combines hands-on training, tailored consultations, and seminars to develop both technical skills and critical understanding within a knowledgeable user community. This conceptual foundation informs DigiCURE's integrated, interdisciplinary methodology, which combines critical heritage studies, digital humanities, conservation science, and archival practice. The infrastructure translates these perspectives into practical workflows through:

- Context-aware metadata modelling co-developed with scholars and heritage professionals;
- Implementation of open standards such as IIIF, GeoJSON, and Linked Open Data to ensure interoperability and long-term usability;
- FAIR-aligned data workflows for sustainable curation, access, and reuse;
- Ethical review protocols for working with sensitive, conflict-affected, or community-based heritage materials;
- AI-assisted analysis that leverages machine and deep learning to support the semi-automatic interpretation of multimodal heritage data – including legacy formats and high-resolution 2D/3D documentation – while ensuring transparency, expert validation, and adherence to ethical standards.

By embedding these principles across all stages – from digitisation and data modelling to AI-supported analysis and spatial exploration – DigiCURE enables a critically informed, context-sensitive, and sustainable engagement with endangered cultural heritage.

Technical Implementation

DigiCURE's technical foundation is a modular, research-driven infrastructure that supports the structured, interoperable, and sustainable management of multimodal cultural heritage data. It consists of two tightly integrated components: *Diana*, the backend data management platform, and *MuM* (Multimodal Map and Viewer), the interactive frontend for visualisation and spatial exploration.

Diana – Data Management Backbone: Diana, a Django-based data management platform, enables place-based modelling, multimodal data integration, and versioned, standards-compliant APIs. It supports multiple data formats and real-time updates and offers user-friendly administrative tools. The platform ensures security, flexibility, and long-term interoperability. Compliance with the FAIR principles [21] is ensured through persistent identifiers, open APIs, and licensing models that facilitate long-term accessibility, interoperability, and reuse. Additionally, the platform offers built-in support for version control, encrypted storage of sensitive data, and real-time data updates with performance optimisation features such as indexing, caching, and asynchronous processing. As such, it is central to DigiCURE's ambition to offer sustainable, research-driven digital heritage environments grounded in scholarly and technical excellence.

MuM – Multimodal Exploration Interface: The MuM frontend, based on Vue3 and Express frameworks, connects to Diana via REST APIs and is organised around four interlinked views. These enable users to filter and examine cultural heritage data across scales – from regional mappings to the detailed examination of visual data. To handle complex visual content, MuM incorporates various open-source libraries for high-resolution 3D models, point clouds, RTI photography, spatial data, and high-resolution imagery. Through its integration with Diana, MuM functions not only as a visualisation tool but also as a conceptual framework for structuring digital documentation in a research-driven way [22].

Collaborative Ecosystem

DigiCURE complements existing digital heritage infrastructures by filling a critical gap and integrating advanced tools for digitisation, spatial and multimodal data management, long-term preservation, and remote scholarly access – functions often fragmented or missing elsewhere.

Relationship to Other Infrastructures

Nationally, DigiCURE aligns with infrastructures like Huminfra, Språkbanken CLARIN, and InfraVis, contributing a specialised platform for high-resolution documentation, digital preservation, and context-sensitive analysis of at-risk heritage. While collaborating with initiatives such as SveDigArk and DARK Lab's Dynamic Collections, DigiCURE offers a distinct approach through tools like MuM and Diana, supporting seamless workflows from field documentation to spatially anchored exploration.

The DigiCURE infrastructure is interoperable by design, and its data models adhere to international standards, allowing for integration and data exchange with existing repositories and platforms. Through GRIDH, DigiCURE collaborates with the national search service for cultural heritage, K-Samsök, and has implemented the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH), which makes DigiCURE a national data provider for heritage data.

At the European level, DigiCURE complements DARIAH and ARIADNEplus [23], and Europeana [24]. By offering archival interoperability, along with tailored tools and training for researchers and others working with endangered heritage under real-world constraints, DigiCURE contributes to EU and UNESCO-linked projects, where GRIDH infrastructure components are already deployed in crisis-driven cultural preservation initiatives. In dialogue with the National Heritage Board (RAÄ) and ARIADNE, DigiCURE will also interface with ECCCH (www.echoes-eccch.eu), which expands the reach of digitised resources and provides sustainable access for researchers.

Taken together, DigiCURE strengthens the national and international infrastructure ecosystem by developing a dynamic multimodal platform, connecting existing nodes, standardising workflows, and providing a previously missing expertise infrastructure and enhanced access for digital cultural heritage preservation.

3.2 Current Collaborations

Since 2019, DigiCURE has supported numerous research projects and institutions through digitisation expertise, development resources, training, and consultation focused on at-risk heritage. While primarily serving researchers at Swedish universities, the infrastructure hosts significant heritage materials from Sweden, the Arctic, Antarctica, the Faroe Islands, Ukraine, Spain, and Italy that are accessed online more than 20.000 times annually. Through these projects, DigiCURE has built strong ties with scholars and institutions who rely on its services or benefit from its state-of-the-art approach.

Notably, DigiCURE serves as a foundational infrastructure for the project DigiCURE:UKRAINE (funded by the Swedish Institute) that provides the Ukrainian heritage sector with expertise, training and a software platform for preserving heritage assets threatened by the war (digicure.dh.gu.se/ukraine/), including the National Museum of the History of Ukraine and the National Library of Ukraine. High-profile research projects, infrastructures, and organisations using DigiCURE include:

- Digital Documentation of Inscriptions in the Saint Sophia Cathedral in Kyiv employs advanced digital documentation techniques to preserve and analyse inscriptions within the historic cathedral, ensuring accessibility for research and conservation [25].
- CHAQ2020 – Cultural Heritage in Antarctica (KTH, Luleå University, University of Gothenburg, the Argentinian Antarctic Institute) utilises DigiCURE's digital imaging and data archiving to safeguard historical sites threatened by extreme environmental conditions [26].
- Swedish Rock Art Research Archives (SHFA) is the world's largest database of rock art documentation and makes use of DigiCUREs accessible tools for analysing high resolution images and 3D models [27].
- Swedish Institute in Rome (SIR), renowned for its Etruscan studies, collaborates with DigiCURE's expertise for advanced digitisation and research-driven platforms for the documentation of inaccessible Etruscan chamber tombs [28].

DigiCURE is collaborating on two additional major initiatives where data collection is planned but has not yet begun. Preserving La Pileta focuses on documenting the Paleolithic art of La Pileta Cave in Benaoján, Spain is an international collaboration between DigiCURE and the Universidad de Sevilla to ensure long-term preservation through the multimodal platform; Pulse of the Weddell Sea (base funding, the Polar Secretariat) is a multidisciplinary research expedition where DigiCURE will provide expertise to secure data and develop methods for digital accessibility of Antarctic cultural heritage.

DigiCURE also contributes to helping Norsk Polarinstitutt's plan for digitally preserving built cultural heritage on Svalbard threatened by climate change and supports RISE (Research Institutes of Sweden) with the tools for annotation of online point cloud visualisations. Outside academia, DigiCURE has, through expertise, supported, among others, the International Council of Museums (ICOM), the National Heritage Board (RAÄ), the Regional Museums, Europa Nostra, the National Museum of the Faroe Islands, Instituto Antártico Argentino, Tecnópolis Buenos Aires, and the National Historical Museums of Sweden.

Availability

Primary academic users of DigiCURE are scholars in fields such as art history, history, archaeology, conservation, heritage management, architecture, religion, ancient languages, and digital humanities, whose needs mirror the current projects. The infrastructure is also a resource for heritage organisations in need of a knowledge boost and expertise to help them digitise endangered monuments, sites, and collections in ways that serve the research community. Notably, computational researchers – including those developing or training models in computer vision, natural language processing, and multimodal AI – will benefit from DigiCURE's structured, high-quality datasets designed for ethically grounded, research-driven machine learning applications.

DigiCURE's online and open-sourced resources are available to anyone and do not require formal requests or fees. Its online platform is designed from the ground up to be accessible, reusable, and openly available to researchers across institutions, disciplines, and national borders, and it facilitates remote access to heritage materials worldwide. Through its modular architecture and standards-based APIs, it allows external users to access, contribute to, and build upon existing datasets via secure, role-based authentication.

References

[1] Harrison, R. 2016. "Anticipating Loss: Rethinking Endangerment in Heritage Futures," International Journal of Heritage Studies, 22(3), 227–242.

[2] Rico, M. (2015). "Safeguarding the Irreplaceable: The UNESCO List of World Heritage in Danger," Journal of Cultural Heritage Management and Sustainable Development, 5(1), 12–25.

[3] Hansson, K., Dahlgren, A. N., & Pargman, T. C. (2022). "Datafication and cultural heritage: Critical perspectives on exhibition and collection practices". Information and Culture, 57(1), 1–5. https://doi.org/10.7560/IC57101

[4] Roiha, J., & Holopainen, M. (2023). "Digging through databases—A case study of Iron Age sites in Finland by generating and analysing keywords". Heritage, 6(8), 5919–5934. https://doi.org/10.3390/heritage6080311

[5] RAÄ (2024) Swedish National Heritage Board's digital strategy.

[6] Paukkonen, N. (2024). "Ten years of photogrammetry and LiDAR: Digital 3D documentation in Finnish archaeology between 2013–2022". Fennoscandia Archaeologica, 41. https://doi.org/10.61258/fa.142220

[7] Westin, J. & Almevik, G. (2024). "Digitising Sensitive Heritage Monuments in Antarctica". ISPRS International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences (1) 2024.

[8] Lindhé, C., J. Ling, C. Liebl, C. Horn, A. Green, M. Peternell, J. Moyano di Carlo, & E. Meijer (in press) "De gåtfulla hällristningarna i Bohuslän – utmaningar och möjligheter i sökandet efter granitens ristare", Makadam förlag.

[9] Maietti, F., Di Giulio, R., Balzani, M., Piaia, E., Medici, M., & Ferrari, F. (2018). "3D data acquisition and modelling of complex heritage buildings". In A. A.V.V. (Ed.), Digital cultural heritage. Springer.

[10] Terras, M. (2022). "Digital humanities and digitised cultural heritage", The Bloomsbury Handbook to the Digital Humanities, Bloomsbury Handbooks, 255-266.

[11] Argyridou, E., Efstathiou, K., Hadjiathanasiou, M., Ioannides, M., Karaoli, A., Karittevli, E., Mateou, M., Panagi, I., & Samara, P. (2023). "Digital holistic documentation of cultural heritage:

Challenges and risks, towards shaping the future". International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, XLVIII-M-2-2023, 95–102. https://doi.org/10.5194/isprs-archives-XLVIII-M-2-2023-95-2023
[12] Bakken Storeide, M. S., George, S., Suneel Sole, A., & Hardeberg, J. Y. (2024). "3D object quality metrics and their differences: How can we evaluate quality of digitization?". In Archiving Conference (pp. 81–87). https://doi.org/10.2352/issn.2168-3204.2024.21.1.17
[13] dh.gu.se/digicure
[14] Björk, L. (2015). How reproductive is a reproduction? : Digital transmission of text-based documents (Publication No. 59) [Doctoral dissertation, Högskolan i Borås]. DiVA. http://urn.kb.se/resolve?urn=urn:nbn:se:hb:diva-881
[15] Westin, J. (2021). "Arosenius Translated. Digitisation as a Rephrasing of Meaning". The Journal Nordic Museology, 31 (1): 40-55.
[16] Drucker, J. Visualization and Interpretation: Humanistic Approaches to Display, MIT Press, 2020.
[17] Svensson, P. (2025). Humane Infrastructures. MIT Press, Cambridge Mass., 2025.
[18] Harrison, R. (2013). Heritage: Critical Approaches. London: Routledge.
[19] Oberbichler, S., E. Boroş, A. Doucet, J. Marjanen, E. Pfanzelter, J. Rautiainen, H. Toivonen & M. Tolonen (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.
[20] Brodén, D., M. Fridlund, C. Lindhé & J. Westin. "Interdisciplinary Digital Project Design", E. Volodina, G. Bouma, D. Dannélls & D. Kokkinakis (eds), The Huminfra Handbook, in press.
[21] Vlachidis, A., Antoniou, A., Bikakis, A., & Terras, M. (2021). "Semantic metadata enrichment and data augmentation of small museum collections following the FAIR principles". In Information and knowledge organisation in digital humanities: Global perspectives (pp. 106–129). https://doi.org/10.4324/9781003131816-6
[22] Westin, J. Bridge, T. & Tomasini, M. (2024). "From the Arctics to Antarctica - A multimodular visualisation of data", Proceedings of the Huminfra Conference (HiC 2024).
[23] www.dariah.eu
[24] www.europeana.eu
[25] saintsophia.dh.gu.se; github 2025–05-13
[26] antarctica.dh.gu.se
[27] shfa.dh.gu.se; github 2025–05-13
[28] etruscan.dh.gu.se; github 2025–05-13