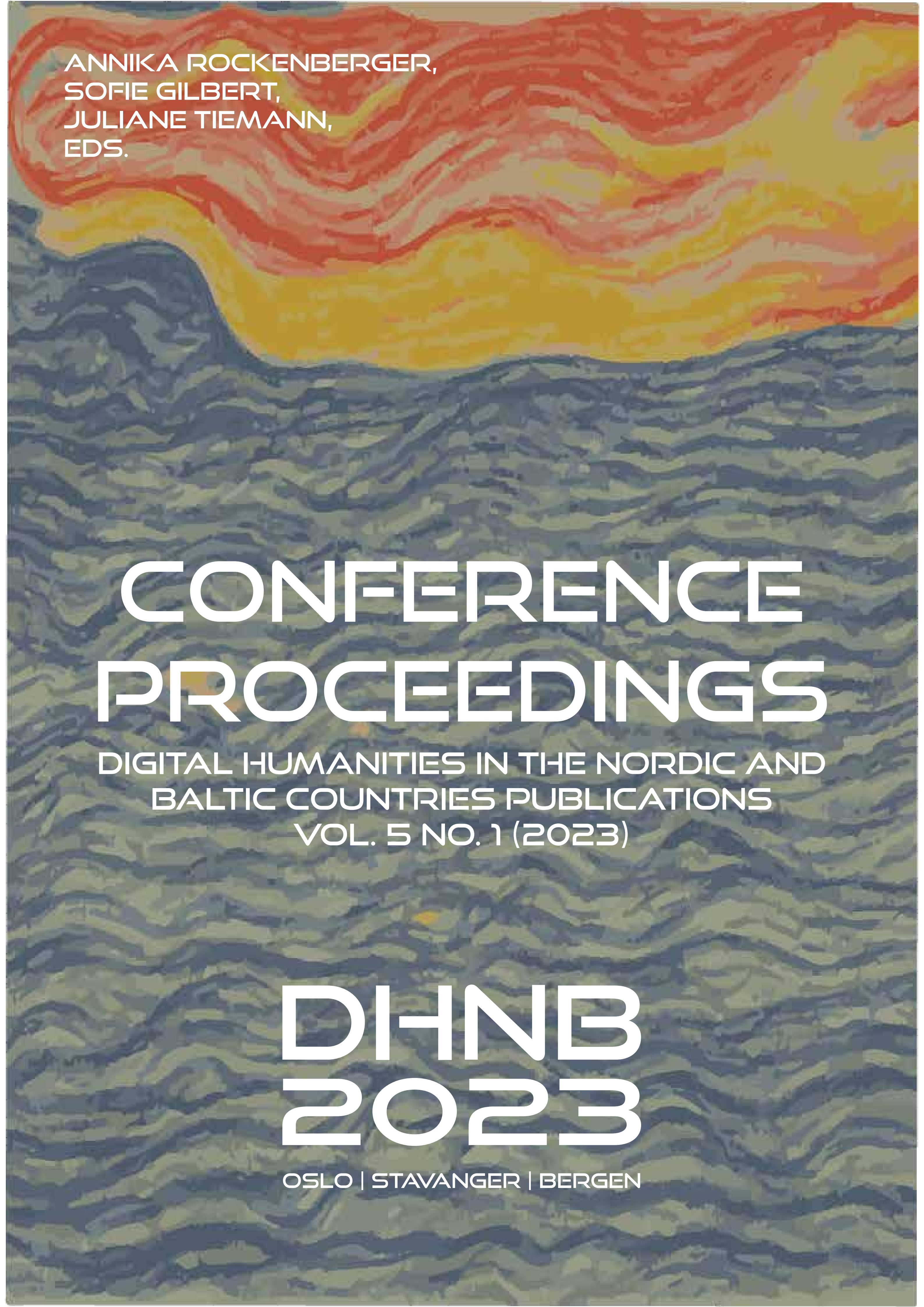


ANNIKA ROCKENBERGER,
SOFIE GILBERT,
JULIANE TIEMANN,
EDS.

CONFERENCE PROCEEDINGS

DIGITAL HUMANITIES IN THE NORDIC AND BALTIC COUNTRIES PUBLICATIONS

VOL. 5 NO. 1 (2023)

DHNB 2023

OSLO | STAVANGER | BERGEN

DHNB2023 Conference Proceedings

DHNB2023
Conference Proceedings

Digital Humanities in the Nordic and Baltic Countries Publications
Vol. 5, No. 1
DHNB 7th Conference

Oslo, Stavanger, Bergen and
online 8-10 March 2023

Edited by Annika Rockenberger,
Juliane Tiemann,
and Sofie Gilbert

UNIVERSITETET
I OSLO

ISBN 978-82-8037-202-4
DOI 10.5617/dhnbpub.10742

Digital Humanities in the Nordic and Baltic Countries Publications
Vol. 5, No. 1 (2023): DHNB2023 Conference Proceedings
https://journals.uio.no/dhnbpub/issue/view/dhnb2023
ISSN 2704-1441

Cover Design by Sofie Gilbert and Annika Rockenberger.

The cover art was created with DreamStudio and the Stable Diffusion latent text-to-image diffusion model using the text prompt "A colour pencil drawing in the style of Edvard Munch of the Nordic and Baltic environment" on 31st August 2022 by Annika Rockenberger. The raw image was then post-processed by Sofie Gilbert, who used it to create the DHNB2023 logos, cover art, and colour scheme.

Published by the University of Oslo Library, Oslo, Norway.

Table of Contents

Editorial

Articles

Editorial. Sustainability: Environment - Community - Data.
The 7th Digital Humanities in the Nordic and Baltic Countries Conference

Annika Rockenberger[1], Sofie Gilbert[1], Juliane Tiemann[2] and Maria Elisa Pierfederici[1]

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

Abstract

Editorial of the seventh annual DHNB Conference Proceedings. DHNB2023 was held online and jointly organized by the University of Oslo Library, The University of Bergen Library, and the Greenhouse Center for Environmental Humanities at the University of Stavanger, Norway from March 8th to 10th, 2023. The conference was shaped by its theme of Sustainability, highlighting and dividing the contributions' sub-themes of Environment, Community, and Data. This preface provides insight into the planning process and the execution of the online conference. In doing so, it offers insight into the choices made regarding the Conference theme and format and visualizations depicting an overview and analysis of conference participants and conference contributions.

Keywords

Digital Humanities, Conference Proceedings, Nordic Countries, Baltic Countries, Participation Analysis, Contribution Analysis

1. Introduction

Foreword from the editors

It is with great pleasure that we present the Proceedings of the Seventh Digital Humanities in the Nordic and Baltic Countries Conference (DHNB2023), held online and hosted by the University of Oslo Library, the University of Bergen Library, and the Greenhouse Center for Environmental Humanities at the University of Stavanger, from March 8th to 10th, 2023. The DHNB2023 conference explored the many facets of "Sustainability in the Digital Humanities," focusing on Environment, Community, and Data.

DHNB2023: Sustainability: Environment - Community - Data, March 8–10, 2023, Oslo, Bergen, Stavanger, Norway

annika.rockenberger@ub.uio.no (A. Rockenberger); sofie.gilbert@ub.uio.no (S. Gilbert); juliane.tiemann@uib.no (J. Tiemann); elisa.pierfederici@ub.uio.no (M. E. Pierfederici)
https://www.ub.uio.no/english/about/people/humsam/annikar/ (A. Rockenberger);
https://www.ub.uio.no/om/ansatte/samdig/apen-forskning/sofiegi/ (S. Gilbert);
https://www.uib.no/en/persons/Juliane.Marie-ThÃĬrÃÍse.Tiemann (J. Tiemann);
https://www.ub.uio.no/om/ansatte/samdig/apen-forskning/mariaepi/ (M. E. Pierfederici)
0000-0001-9515-8262 (A. Rockenberger); 0000-0002-3180-6010 (S. Gilbert); 0000-0003-0893-701X (J. Tiemann); 0000-0002-6882-6576 (M. E. Pierfederici)

DHNB Publications, DHNB2023 Conference Proceedings, https://journals.uio.no/dhnbpub/issue/view/875

At the 2022 Annual Members' Meeting (held in conjunction with DHNB2022), we first announced the plans for the DHNB 2023 conference. We had just received confirmation from all partners that they were willing to collaborate and organize a fully online event. A condition for our partnership was to agree on a topic that would span widely across the Digital Humanities and Social Sciences, be of concrete relevance to our diverse institutions, attract many actors in the Nordic and Baltic countries, and serve as an essential and timely contribution to the current discourse in academe. We soon found that 'Sustainability' was that topic, with the sub-topics representing particular focus areas of our three partnering institutions: The Greenhouse Center for Environmental Humanities with its expertise in the sustainability aspect 'Environment'; The University of Oslo Library and its Digital Scholarship Center's dedication to building and sustaining communities, and the University of Bergen Library's decade-long investment in humanities and social sciences data collection, curation, and dissemination.

These proceedings cover the three sustainability aspects guiding the conference under the three thematic tracks – *Environment, Community,* and *Data* – and include the full versions of long papers submitted after the conference.

Environment

The Digital Humanities do not stand at a distance from the environmental challenges facing the planet. In 2015, Bethany Nowviske challenged DH scholars and practitioners to consider the field's role in the Anthropocene (Nowviske, 2015). What responsibilities do we have as the world around us burns, dries, drowns, and changes before our eyes? How do DH projects and practices depend on unsustainable systems and mindsets? How do the unequal consequences of environmental challenges influence what research is conducted in DH, and who can contribute? How can the field of DH contribute to a more sustainable world?

Community

Since its inception, Digital Humanities has been a community-driven effort. We can see this not least in the many regional and linguistic organizations all over the globe. The Digital Humanities have been described as grassroots communities, sprouting from small local research groups or gathering around digital research support centers and labs at universities and libraries. DHNB is a young and prosperous community spanning eight countries and speaking many languages. However, is it a sustainable one? Sustainable communities are places where people of diverse backgrounds and perspectives feel welcome and safe, where every group and member has a say in decision-making, and where intellectual prosperity is shared. What does this mean for DHNB now and in the future? Lastly, how can the community continue to be successful together?

Data

The primary source material for humanists has many data formats; meanwhile, research is becoming increasingly digital and, in many cases, is only available in digital formats. As increasing digitization leads to a large volume of data, Digital Humanities must implement affordable ways to access, store, and archive these data. The efforts of doing so can be seen in developing large data repositories, both collectively and within specialized fields. When it comes to making the collected data of repositories, but also of single examinations, accessible – and as such also visible – Open Data/Open Science has become a well-known term and a requirement in many funding evaluations. Nevertheless, what does this mean in terms of sustainability? How does the growing amount of digital data available for research within Digital Humanities go together regarding long-term storage, communal access, and the restrictions of sensible data? What aspects of collaborative software development concerning future accessibility could help with the environmental footprint of this data volume?

2. Planning DHNB2023

The planning of the DHNB2023 conference started early in 2022 and was, as the conference itself, done online. Even though post-corona times have prepared us well for organizing and realizing digital events, conference planning is challenging. And as accustomed as we all have become to participating in conferences in front of a screen, we still wanted to give participants who were in Norway at the time of the conference the possibility to celebrate this event on-site with the local organizers, even if it was just for one evening. The planning thus ended up being one for an online conference, while the three keynotes and receptions were held on-site in Stavanger, Bergen, and Oslo. We were happy to greet everyone who joined us for these local events, which were live-streamed.

During the conference, we were lucky enough that many colleagues within the digital humanities community were willing to chair the various parallel sessions during the three days. With local Zoom hosts from Oslo, the sessions were divided between a host and a chair, smoothly implementing the sessions with each two to three papers. We divided contributions into long papers and show-and-tell presentations. While long papers were presented live, followed by a discussion with the participants, the show-and-tell presentations were pre-recorded and ran in reels throughout the conference. They were also made available online on the DHNB Youtube channel for DHNB2023. The three keynotes reflected the three aspects of Sustainability chosen for this conference (Environment, Community, and Data) and were held as hybrid events.

Lisa Swanstrom, associate professor at the Department of English at the University of Utah, presented the thematic track of environment and talked about "Forecasting sustainability. Speculative ecologies at work in DH, EH, and AI". She identified a familiar yet misleading story circulating within AI to outline an alternative aesthetic genealogy. Swanstrom further discussed the importance of re-framing AI within literary studies and the Digital Humanities in a way that confronted its statistical underpinnings. She offered a different approach to Natural Language Processing within literary scholarship committed to a broad, material, and environmentally responsible concept of intelligence—artificial or otherwise. The Greenhouse hosted this keynote at the University of Stavanger.

Malvika Sharan, senior researcher at the Alan Turing Institute, was the representative for the thematic track of community and talked about "Open Science for enabling reproducible, ethical and collaborative research." She discussed open science as a framework to ensure others can easily access, openly examine, and build upon research components. In her talk, she shared best practices researchers should integrate from the start of their projects to maintain the highest reproducible and ethical standards to guarantee that their research is easy to reuse and reproduce at all stages of development. The University of Oslo Library hosted the keynote, with a reception beforehand.

Scott Rettberg, professor of digital culture at the Department for Linguistic, Literary, and Aesthetic Studies at the University of Bergen, was the representative for the thematic track of data and talked about "Building a sustainable research infrastructure. The ELMCIP electronic literature knowledge base". He described the process and challenges of developing and maintaining the project "Electronic Literature as a Model of Creativity and Innovation in Practice" until today. He also discussed the changes to the database planned for new uses in the Center for Digital Narrative, a Norwegian Center of Research Excellence that is launching in August 2023. The University of Bergen Library hosted the keynote, with a reception beforehand.

3. DHNB2023 in numbers

The initial Call for Submissions received 85 proposals for the formats of Long Papers, Show-and-Tell Presentations, Panels, and Workshops. The Program Committee selected 68 submissions for the conference after a single-anonymized peer-review process: 44 Long Papers, 20 Show-and-Tell Presentations, and six workshops. The submissions spanned various topics within Digital Humanities, from 3D modelling to web research. We have gathered the presentations for the program into three thematic tracks – *Environment, Community,* and *Data* – and two open tracks, focusing on specific DH methods or research objects. The keynotes featured each day's central theme and concluded the day.

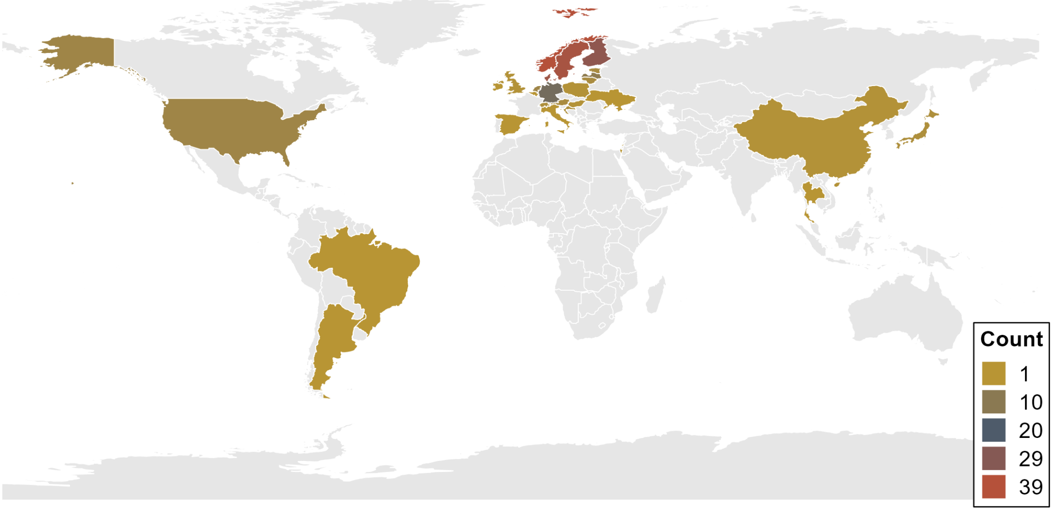


Figure 1: World map of conference participant locations

DHNB conferences attract an international academic community. DHNB2023 received proposals from 19 countries: Australia, Austria, Belgium, Bulgaria, China, Denmark, Great Britain, Estonia, Finland, Germany, Israel, Japan, Latvia, the Netherlands, Norway, Poland, Sweden, Switzerland, and the United States of America.

The benefits of an online conference were clear when examining our participants' and authors' locations. Two hundred fourteen participants participated in the event worldwide, spanning 28 countries, as shown in Figure 1 and Figure 2. Norway had the highest number of participants at 39, followed by Sweden and Denmark with 36 and 34 participants, respectively.

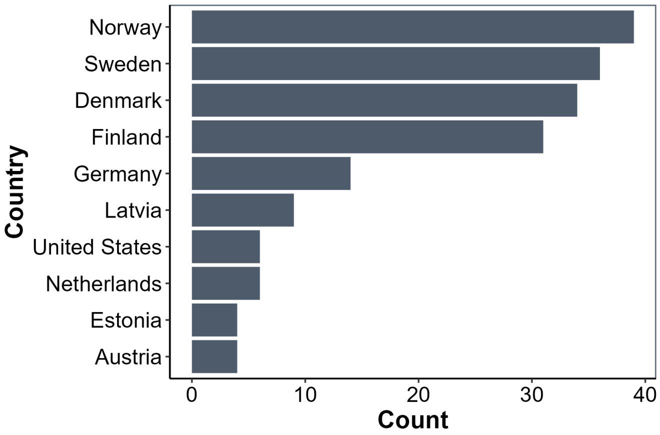


Figure 2: Histogram depicting the ten most common participant locations

Most authors' institutions are in the Nordic and Baltic countries. Aalto University in Finland had 16 authors contribute, while Aarhus University in Denmark had eight, and Uppsala University in Sweden had five. Meanwhile, the following institutions had four authors contribute: Austrian Academy of Sciences, Austria; UiT Arctic University of Norway, Norway; University of Gothenburg, Sweden; University of Helsinki, Finland; and Vrije Universiteit Amsterdam, The Netherlands, as shown in Figure 3.

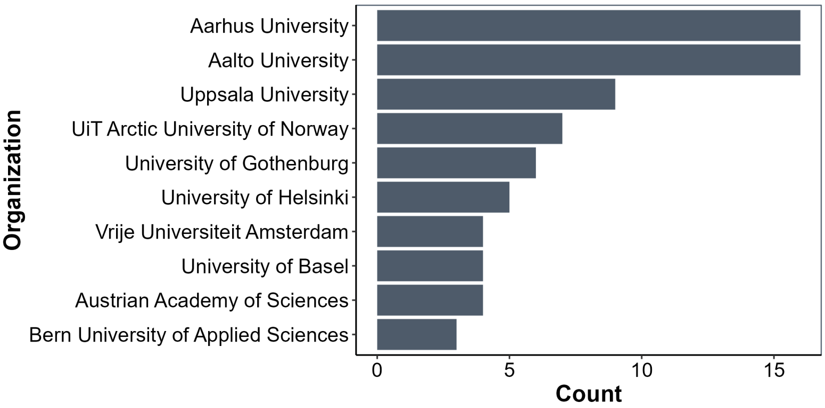


Figure 3: Histogram depicting the ten most common author affiliations

4. Conference Proceedings

For the proceedings of DHNB2023, we wanted to produce the proceedings shortly after the conference concluded. Authors with accepted papers were given information about the publishing plan at the end of 2022, including an intended timeline for the submissions we planned to publish with CEUR workshop series, which had been the DHNB conference outlet since 2018. Regrettably, in July 2023, our Digital Humanities proceedings were denied by CEUR as they no longer fit their newly revised topic requirements. The DHNB Board and the University of Oslo Library stepped up and decided to create a conference and workshop proceedings outlet that would meet our needs and those of future conference organizers by establishing DHNB Publications as part of the University of Oslo's FRITT Diamond Open Access Platform for serial publications. We publish all articles under a CC-BY 4.0 Creative Commons Attribution 4.0 International license, and all authors retain the copyright to their work. Each article will be assigned a Digital Object Identifier (DOI) and enhanced with bibliographic metadata.

The articles of these proceedings take up 10–15 pages each. All articles underwent single-anonymized peer review by at least two reviewers from the digital humanities community. After a revision round, the final versions are presented in these proceedings.

5. Review Process

For DHNB2023, the organizers and the Program Committee decided on a different approach for the reviews than we so far followed for previous conferences. For the first time in the history of the DHNB conference, reviewers received instructions about the concrete parameters for evaluation and how to provide feedback to authors. We first had a round of single-anonymized reviews of abstracts. These were done by the members of the program committee, who reviewed 20+ submissions each. The quality of submissions was generally high, and we were happy to accept all workshop proposals. The acceptance rate of long papers was 65%; three long papers and one panel were suggested to be turned into show-and-tell presentations. We accepted 80% of show-and-tell presentations. We then invited accepted long papers to submit a full version for the conference proceedings. We also sent a call for reviewers to the community, to which we received 43 responses. These peers received instructions on the parameters for evaluating full papers and the feedback to authors. Of the 30 submissions we received, we were able to accept and publish 25 in these proceedings.

The three most common topics chosen by authors were computational text and literary analysis, computational text processing, and cultural heritage collections. However, the number of topics selected by DHNB authors varied and covered a broad range of topics within the theme, as shown in Figure 4 below.

We proudly present these high-quality papers to the Nordic, Baltic, and international Digital Humanities community, and on this note would like to thank the community reviewers for their excellent work: Emma Aadland, Maria Akritidou, Nikolay Atanasov, Daniel Brodén, Bastiaan Bruinsma, Gimena del Rio Riande, Senka Drobac, Elena Fernandez Fernandez, Mats Fridlund, Béatrice Gauvain, Elizabete Grinblate, Raphaela Heil, Alíz Horváth, William Illsley, Heidi Jauhiainen, Tommi Jauhiainen, Mica Jorgenson, Maija Kāle, Ernesta Kazakėnaitė, Anders Klindt

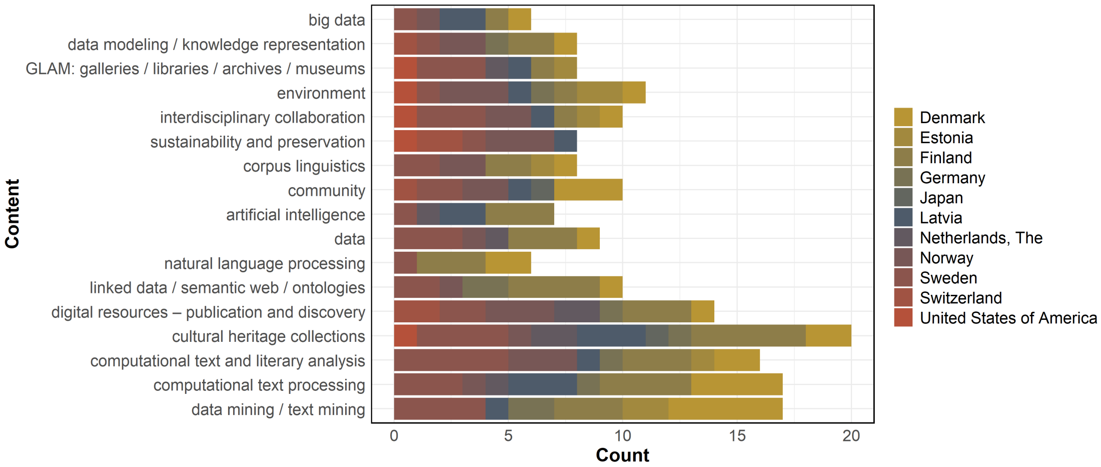


Figure 4: Most popular topics by country

Myrvoll, Ross Deans Kristensen-McLachlan, Nataliia Lazebna, Ying-Hsang Liu, Kateryna Lut, Elisabeth Maria Magin, Daniele Metilli, Liisa Näpärä, Seraina Nett, Kristoffer Nielbo, Sebastian Lundsteen Nielsen, Dalia Ortiz Pablo, Federico Pianzola, Anna Kristiina Ristilä, Torsten Roeder, John Charles Ryan, Maria Skeppstedt, Karina Šķirmante, Pål Steiner, Jon Carlstedt Tønnessen, Jurgita Vaičenonienė, and Andrew Wareham.

6. Outlook

DHNB will again hold its annual conference on-site after three years of online and hybrid conferences. From the 27th to 31st of May 2024, the community will gather in Reykjavík, Iceland, where the Centre for Digital Humanities and Arts (CDHA) / Miðstöð stafrænna hugvísinda og lista (MSHL) at the University of Iceland will organize the 8th conference. CDHA is a collaboration between 11 institutions: Árni Magnússon Institute for Icelandic Studies, Iceland University of the Arts, Icelandic National Broadcasting Service, National and University Library of Iceland, National Archives of Iceland, National Gallery of Iceland, National Museum, Rekstrafélag Sarps, Reykjavík Art Museum, University of Iceland – School of Humanities and University of Reykjavík. It is the first time a DHNB event will be held in Iceland, one of our most remote regions.

We are also in the fortunate situation to already have a host for the DHNB2025 conference. This time, Estonia will be the hosting country, with the Estonian Literary Museum, the Estonian National Museum, the University of Tartu, and the Estonian Society of Digital Humanities jointly organizing the 2025 conference.

References

[1] Nowviskie, Bethany (2015). Digital Humanities in the Anthropocene. Digital Scholarship in the Humanities, 30(1), 4-15. https://doi.org/10.1093/llc/fqv015

Benign Structures. The Worldview of Danish National Poet, Pastor, and Politician N.F.S. Grundtvig (1783-1872)

Katrine F. Baunvig[1,2,3], Kristoffer Nielbo[2,3]

[1] *Center for Grundtvig Studies (Aarhus University), Jens Chr. Skous Vej 3, 8000 Aarhus C, Denmark*
[2] *Center for Humanities Computing, Jens Chr. Skous Vej 4, Building 1483, 4th floor, 8000 Aarhus C, Denmark*
[3] *Aarhus University, Nordre Ringgade 1, 8000 Aarhus C, Denmark*

Abstract

This paper presents a study of the central Danish 19th-century figure N.F.S. Grundtvig (1783-1872) and his worldview. More precisely, by way of simple neural word embeddings we seek to plot Grundtvig's worldview in attempt to tease out factors explaining for his cultural and religious success as a poet. We arrive at the conclusion that one factor for Grundtvig's relative success could be found in his explicitly positive worldview fit for modern conditions.

Keywords

Danish Cultural Heritage, Worldview, Neural Word Embeddings, N.F.S. Grundtvig, Grundtvig's Works, Grundtvigs Værker

1. Introduction: N.F.S. Grundtvig. A Cultural and Religious 'Saint'

In Denmark N.F.S. Grundtvig (1783-1872) plays the dual role of Church Father and Founding Father. In public discourse the 19th-century poet, pastor, historian, antiquarian, educator, and politician is regarded as one of the most (if not *the*) central figure in the Danish nation building process, as well as in the reformation or adaptation of Christianity to modern conditions: In short, he is regarded little short of a cultural and a religious saint.[1] In scholarly literature it is widely acknowledged that Grundtvig sought to stimulate the process of assembling a collective Danish emotional consciousness based on a horizontal-contemporary axis incorporating the

DHNB2023 | Sustainability: Environment - Community - Data. The 7th Digital Humanities in the Nordic and Baltic Countries Conference. Oslo – Stavanger – Bergen, Norway. March 8–10, 2023.

✉ baunvig@cas.au.dj (K. F. Baunvig); kln@cas.au.dk (K. Nielbo)
🌐 https://grundtvigcenteret.dk/ (K. F. Baunvig); https://au.dk/kln@cas.au.dk (K. Nielbo)
0000-0001-6706-6694 (K. F. Baunvig); 0000-0002-5116-5070 (K. Nielbo)

DHNB Publications, DHNB2023 Conference Proceedings https://journals.uio.no/dhnbpub/issue/view/875

[1] This 'sainthood' is not a banal, cosmetic analogy. Reverence and quasi-ritual structures have been built around Grundtvig as a Great Dead [1]). Grundtvig has a cathedral named after him: the Copenhagen Grundtvigs Kirke; every year his birthday (almost coinciding with the day of his death) is celebrated in Grundtvig-relevant institutions; one such celebration entails the opening of his crypt at the small cemetery Clara's Kirkegård on the outskirts of the Sealandic town of Køge. Moreover, Grundtvig's Death (Grundtvigs død) is a commodity – at least it is a fairly recent title in a popular book series by Aarhus University Press written by Grundtvig scholar Jes Fabricius Møller (2019) [2].

different strata within the socially heterogeneous “Folk” [the People], and on a vertical-historical axis connecting present-day Danes with forefathers and legendary characters. In social historian Benedict Anderson’s well-known phrasing, the emotional fabric intended by this attempted interlacing was an ‘imagined community’. Nowadays, Grundtvig’s cultural imprints are acknowledged by most Danes: “N.F.S. Grundtvig founded Danish democracy”; “N.F.S. Grundtvig established the Evangelical-Lutheran Church in Denmark (*folkekirken*)”; “N.F.S. Grundtvig is the founder of the Danish school system”; “N.F.S. Grundtvig revived the pre-Christian Nordic tradition”; “N.F.S. Grundtvig is the most important writer of Christian hymns in Denmark”. These are surprisingly recurrent statements in Danish public media, deeming his intellectual activity more culturally important than the work of his world-famous contemporaries Søren Kierkegaard (1813–1855) and Hans Christian Andersen (1805–1875).

2. N.F.S. Grundtvig: Highly Popular Hymnist

In addition to such statements, the public sphere – not least the political niches – are overflooded with quotes from his many well-known hymns and songs; due to their mere number – Grundtvig wrote no less than 1,600 very long hymns and songs –, they are conveniently versatile in political ideology and can be tweaked or cherry-picked into supporting any given agenda. But the main point here is, that politicians or public characters of different breeds still find it relevant to evoke Grundtvig in current-day debates. Seemingly, he can bolster any given person or any given argument and add a certain empathic ethos to and unyielding air around them. This situation is a result of the stable exposure to Grundtvig’s hymns and songs that Danes experience, we suggest. For a signature feature of Danish culture is a proneness to communal singing – in kindergardens, in elementary schools, at the workplace, in Folk High Schools, in civil societal associations, in church, in various private settings. And Grundtvig is a dominant hymnist and secular song writer. In illustration: he has written no less than a third of the close to 800 hymns in the 2003 Danish official hymnal.

3. Why did Grundtvig become so popular?

In this paper, we have set out to explore reasons for Grundtvig’s poetic success. He was by no means the only poet of the 19th-century. A long line of poets, hymnists, and song writers joined him in his endeavor to improve and boost Danish literature during this historically critical period. Furthermore, a line of hymnic projects in Denmark preceded him. A high-profile but remarkably unsuccessful candidate who predated Grundtvig would be Bishop N.E. Balle (1744-1816). He commissioned a hymnal in the late 18th century that would meet the taste of an enlightened bourgeoise elite. It was published in 1798 as *Evangelisk-kristelig Psalmebog til Brug ved Kirke- og Huus-Andagt [Evangelical-Christian Hymnal for Church and Private Use]* . It came in the wake of the national debate on the quality of the hymnbook, which included contributions from the weekly periodical, *Jesus and Reason*, edited by the theologian, Otto Horrebow (1769-1823). His purpose was to present Christianity as “the only, eternal, and true religion of Reason” and it should take as its starting-point the development of the individual’s feelings of happiness. Such was the effect of this impulse that theologians and civil servants began to look around for hymns that could "express happy feelings" [3], and a committee was set up to produce a hymnbook that appealed more to those who were critical of their contemporary Church as well as to an increasing individualism among the bourgeoisie. The new hymnbook was full of deism and Enlightenment Christianity, purified of the old myths, superstitions, and metaphysics. Unsurprisingly, there was opposition from the rural parishes, and a complaint that "the Devil has disappeared without trace from the hymnal" (p. 165) [3] – and the hymnal was deemed unsatisfactory, bland, and boring: it did not catch on. N.E. Balle's *Evangelisk-kristelig Psalmebog til Brug ved Kirke- og Huus-Andagt* was to a great extent an attempt to replace the so-called *Kingo's Salmebog [Kingo's Hymnal]* of 1699. This was a very popular hymnbook that contained many but not solely hymns written by Danish priest, poet, and Bishop Thomas Kingo (1634-1703). The problem with the Kingo hymns was that they were not enlightenment compliant and that they did not invite for meditations over how to achieve individual and national happiness. Quite the contrary, they promoted an early modern morbid *memento mori* ethos. Thomas Kingo's concern with the physical factors of death is best illustrated by his hymn from 1674, 'The sun is on the wane', no. 761 in the current Danish hymnal, which contains the following lines: "Dark grave indeed/with worms at feed/is our last fate on earth…/Go, sack of worms, and sleep,/for God will wake you in the morn/to everlasting life." As Danish scholar of religion Hans J. Lundager Jensen satires, Kingo saw "cadavers, where one might naively see living bodies" (p. 25) [4]. Grundtvig agreed with Balle that the morbid aspects of the still popular Kingo hymns were inappropriate and outdated. But he did not think that *Evangelisk-kristelig Psalmebog til Brug ved Kirke- og Huus-Andagt* had solved the problem – the lyrical quality was to poor and the erasure of the long line of Christian fantastic features (monsters, miracles etc.) straight-out demonstrated poor judgement, in his opinion. Conversely, Grundtvig praised the Kingo hymns for their lyrical quality but had problems with the dark outlook upon life that they promoted. And it seems that Grundtvig's evaluation was correct. At least he eclipsed both Balle and Kingo; and, getting a bit ahead of our analysis, he did so by offering a different worldview – one more fit or relevant.

4. Structures of a Worldview

No other poet comes close to Grundtvig's position in Denmark. This fact might obviously have something to do with aesthetic quality of the lyrics as well as of the melodic side of the song; it might also have something to do with authorship brand that can be but does not have to be affected by aesthetic features. In this study, we leave such aspects aside and focus on the worldview representation emerging from the lyrics. Our premise is that hymns and songs that enters a given cultural marked and quickly dominates it, such as Grundtvig's did, must be offering something in demand. They must be the answer to some need. And being as Grundtvig was by no means the only agent on the market, it would be reasonable to assume that there is something in the lyrics, and in his authorship as such, that people found attractive or at least relevant. Our hypothesis is that in a period of significant cultural, political, and religious change this something is to be sought out in the general outlook on life – in the worldview conveyed in the texts. By 'worldview' we mean to invoke American Social Anthropologist Clifford Geertz's definition. That is, a sort of 'background knowledge' of the world, and possible adjoining spheres, and the mechanisms driving life herein. That is: basic ideas of ‘the ontic’ (of what and why things and creatures exist) and under which conditions they do so. In short, we wish to map out Grundtvig’s understanding of the world. We do this on the basis of a data set consisting of Grundtvig’s published writings.

5. The Grundtvig Data

The dataset, *Grundtvig’s Works*, represents the total number of works published in Grundtvig’s lifetime (N = 1073). The first work published is dated 1804, the last one 1872. This material has been OCR prepared and is being furnished with XML markup by the staff of Center for Grundtvig Studies, Aarhus University, following TEI guidelines. Currently 54% of the material is fully annotated. The process of enrichment is, however, ongoing: the project’s scheduled completion date is 2030. The data set comprises approximately 37 K pages, has a median document size of four pages, and contains 4MM word-tokens distributed over 115 K word-types. The data for the current study are available at: https://github.com/centre-for-humanities-computing/grundtvig-data. Furthermore, we have developed a custom XML parser available to facilitate third-party data exploration. The parser is available at: https://github.com/centre-for-humanities-computing/GrundtvigParser The way we went about tweaking out Grundtvig’s worldview from this material was by way of the well-known text analysis strategy: simple neural word embeddings.

6. Simple Neural Word Embeddings

Neural word embeddings are learned low-dimensional representations of discrete data as dense arrays. Condensing a high dimensional space, such as the *Grundtvig’s Works* vocabulary, into a denser one, the procedure allows for visualizations of, e.g., a given term’s ‘semantic habitat’. Put in another way, it allows for teasing out the associative structures by which a given word (or a given compound of words) is nested within a corpus. As is well known, “the map is not the territory” [5], but as is similarly well-known functional maps do, nevertheless, retain certain relevant properties of the territory it seeks to represent. This is also the case with neural word embeddings. While reducing the complexity of a high dimensional semantic space, relevant properties from the original space is preserved. Further, they are preserved in a way affording them to be conveyed as geometrical ratios [6]. This makes it possible to display semantic structures with an enhanced clarity. The structural representations, or the *embeddings*, are ‘learned’. That is, they are the product of a training process over the data in an artificial neural network. A neural network uses a so-called auxiliary control task to iteratively learn the best possible embedding of a word (ex. by minimizing prediction error). In this particular study, we deployed a so-called hierarchical softmax technique to train the neural network [7]. This choice was partly based on the fact that this approach has proved itself advantageous for dealing with infrequent words [8]. In order to explore the associative structure between given terms in the Grundtvig data, we constructed an algorithm that utilize geometric distances between neural embeddings to create (seeded) hierarchical semantic graphs. This algorithm generated graphs by computing the distance between, on the one hand, a given seed term (or compounds of seed terms) and the corpus lexicon *in toto*, on the other, using the inverse trigonometric arccosine function. The series of *seed terms*, to evoke Grundtvig's worldview, we quite simply decided to be the dominating ontic domains in his writing: HEAVEN, EARTH, and HELL. For each seed, the algorithm excerpted a pre-set number of *primary associations.* These are the terms with the shortest distance to the seed terms. For each of the *primary association*-terms the algorithm extracted a pre-set number of association; these associations to the primary associations are taken as *secondary associations* to the seed term. For this study the number of primary and secondary associations, respectively, was 10. The next step was to compute the distance between the respective categories of terms (seeds, primary associations, secondary associations); subsequently the bulk of terms were connected based on their distance under a given threshold estimated from the distance variance structure. At the final stage, semantic clusters (or 'communities') were unearthed by way of a so-called greedy optimization method: the Louvain method [9]. Visually, the graphs render terms as nodes and thresholded distances as edges. For reading purposes, node colour was chosen to specify the given term's semantic group; further, the UPPER CASE was used to distinguish seeds and primary associations from secondary associations rendered in the lower case.

7. Analysis: The Benign Structures of Grundtvig's Semantic Worldview

What we see in the visualization of Grundtvig's semantic worldview in Figure 1 below is that the semantic tissue binding together the spheres of Himmel [Heaven], Jord [Earth], and helvede [Hell], is the radiant, divine Soel [Sun] – embraced or enclosed by a garden-like zone. The structure holding together Grundtvig's conception of the universe is, in other words, remarkably 'positive', warm, fertile, and benign. Further, the central position of the divine sun signals a strong and sustainable integration of the godly and the human sphere. Let us delve into this aspect of Grundtvig's writings. 'Hell' enjoys no such connection. In fact, ´Hell', and the semantic neighbor 'Death', are strikingly isolated in Grundtvig's worldview – pushed out into the far and loosely connected edges. 'Heaven' on the other hand seems to mingle and intertwine with the earthly sphere. The incitement of 'Heaven' and the suppression of 'Hell' is not a trivial trait. A so-called world-denying outlook – stressing the malignant, hellish qualities of earth and life on it – formed within the world religions in the centuries leading up to the Common Era's beginning [10]. It is a classical argument that this development undergirded anthropocentrism and human exploitation of natural resources, e.g. [11]. In the Protestant traditions the world-denying outlook fared particularly well within the variety of awakenings convulsing in Europe and the Americas in the 17th, 18th, and 19th century. But evidently not in Grundtvig's mindset – a mindset imprinted in his successful poetry. Grundtvig provided a positive, world-affirming outlook relevant for 19th-, 20th-, and 21st-century Danes experiencing increasing levels of comfort and to whom 'blessing' was and is a more plausible semantic framework than 'damnation'. In other words, the embedding of Heaven, Earth, and Hell points to the possible explanation for Grundtvig's poetic success: He wrote lyrics imbued with a worldview fit for or culturally sustainable in modernity.

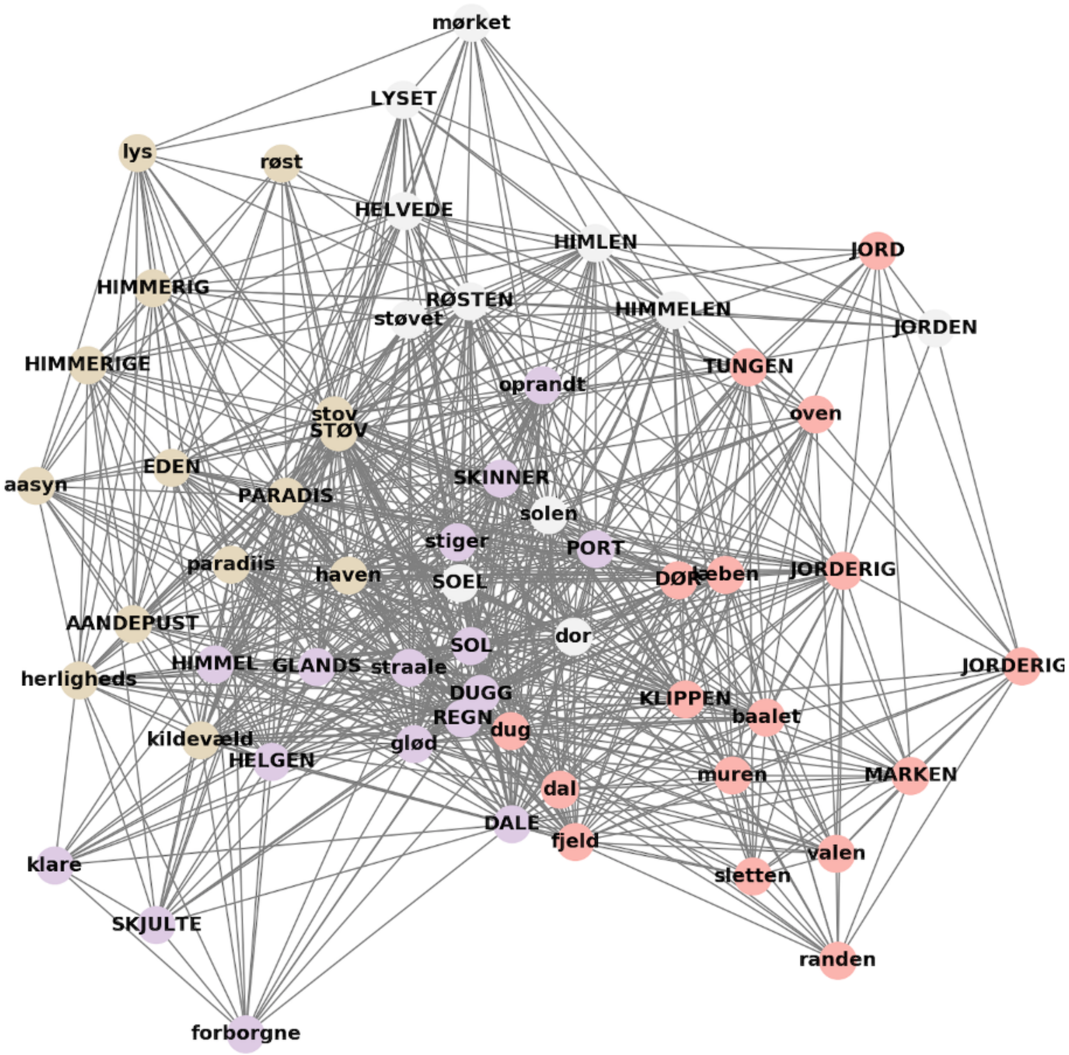


Figure 1: The Semantic Worldview of N.F.S. Grundtvig. The center of the graph conjoins a group of 'solar' terms [Soel (sun), SKINNER (shines), stiger (rises), straale (beem/shines), glød (glow)] girded by zone of fertility-horticultural or Eden-like terms [haven (the garden), PARADIS (paradise), EDEN (Eden), DUGG (dew), regn (rain)]. To the left, this zone seamlessly morphs into a cluster of heavenly terms [HIMMERIGE (Heavenly Kingdom), herligheds (glory's), lys (light), kildevæld (fountain)]. To the right, it gradually goes through terms indicating correspondence and contact [PORT (gate), DØR (doorway)] becomes a group of earthly terms [JORDERIGE (earthly kingdom), JORD (earth), JORDEN (the earth), MARKEN (the field)]. In the upper outskirts of the graph, HELVEDE (hell) lingers loosely connected to the contrast term HIMLEN (heaven) and to it's semantic partner mørket (the darkness).

8. Discussion: The Benign Structures in Grundtvig's Writings

Grundtvig was more interested in life before, rather than after, death. He represented a Christianity that acknowledged life on earth as being fundamentally good, in contrast to the strongly 'tantric' focus on suffering and death which had exerted a large and varied influence on western Christianity since its early days of the Christ movement. Grundtvig insisted that even Jesus' life and violent death should be seen as a story of "the happiest human life" (p. 227) [12]. Grundtvig's 'jubilant' acceptance of life on earth is apparent in his hymns and in the embedding of Figure 1. The visualization echoes the hymns full of earthly joys and beings: skylarks, flowers, linden trees, beechwoods, rushes, dew, beaches, clouds, thunder, lightning, stars and suns, the morning sunrise and the evening sunset, noonday and midnight hours, the wind in the trees, tongues and voices, songs and hearts, hands, smiles, eyes etc. In other words, Grundtvig paints situations and positions that singers of his hymns could be expected to find themselves in, and which he never encourages them to leave. On the contrary, his poetry is a "humble recognition" (p. 71) [13] and a respectful celebration of this earthly life as God's creation – a "song-sacrifice" that rises from "our lips to Heaven" (p.4) [14]. On this point Grundtvig's hymns have clear links to the archaic songs of praise in the Old Testament – 'archaic' in the sense used by Robert N. Bellah in his theories of religion and cultural evolution connoting world-affirmative logics and pre-occupations with themes of fertility. It is of further interest that in contrast to certain theological thinking from the 18th century onwards which appeared to undermine the Old Testament source-material, Grundtvig saw no qualitative difference between the Old and the New Testament (p. 66 et passim) [15]. This is in line with his view that priority should be given to the living church congregation that he thought imbued with the Holy Spirit rather than to the 'dead letters' of the Bible, whether they be in the Old or the New Testament. The Gospel preceded the gospels, he argued. In similar vein, the point of gravity in Grundtvig's hymnwriting is the Earth below rather than Heaven above – for Grundtvig the earth "is significant in itself" (p.36) [16]. Grundtvig even conceives life after death in earthly terms, often as a pleasant garden recalling summer holiday memories [13]. Time and again he employs the term "God's garden" to depict the afterlife, not least in his existential hymns about death and in his eschatological hymns about the time after the second coming of Christ. The garden as an image of life after death has a long Christian history: McDannell and Lang demonstrate the role played by Eden and the rose-garden in the Middle Ages and in Renaissance ideas of life after death (p. 70 et passim; 112-124) [17]. At the same time, we hear in Grundtvig's application of the image in his famous hymn 'Then the wilderness shall bloom/ like a rosy bower' an echo of Martin Luther's vision of the restoration and cleansing of the earth after the Day of Judgement, when Heaven and earth will be a new paradise (ibid., 152) [17]. "Our garden round with happiness is bordered", Grundtvig stated in one of his many 'garden' poems. His focus on the garden is later expanded into a special interest in horticulture among Grundtvigians in the second half of the 19th century. In Grundtvig's thinking the earthly garden as a vision of paradise overshadows the celestial light of Heaven that had been its main rival in Christian imagery since scholasticism (p. 80 et passim) [17]. There is plenty of 'light' in Grundtvig's hymns, but his interest in the powerful, holy sun lies more in the fact that it beams down on Earth. Moreover, for Grundtvig the special status of the Earth has a parallel in the decimated status of Hell [18] mentioned above. He does not often mention Hell, but when he does, it is in line with his dedication to ‘this side of the grave’, and in contrast to the view of Hell in the *Evangelisk-kristelig Psalmebog til Brug ved Kirke- og Huus-Andagt* of 1798, where the divine sphere dwindles even as the life sphere of the sinner is emphasised. In Grundtvig’s hymns the earthly world is given meaning by the very fact that it is divinely created for humans: “without You/ empty is the very earth”, he writes in the hymn, “Come, God Holy Spirit, come”(1837), implying that the Earth is in fact not empty. It is swarming with life, with beings, with content, and occasionally God reaches out his “hand from above!/Its radiance reaches our transient clay,” thus emphasizing the link between the two spheres, Grundtvig observes.

Many writers have rightly pointed out the close connection between the earthly-human and the heavenly-divine in Grundtvig’s thinking. A good example is his reworking of Jacob’s dream at Bethel in Gen 28: 10-22, since Jacob was also “one of Grundtvig’s key figures” (p. 125) [19]. This was a story that he returned to time and again. In all its brevity it tells how one evening Jacob on a journey in the desert lies down to rest with a stone as his pillow; he dreams of God’s angels going up and down a stairway, at the top of which is God, who promises to watch over Jacob and his people. Jacob wakes up in fear: “How awesome is this place! This is none other than the house of God; this is the gate of heaven.” Scholars, most prominently perhaps German scholar of religion Rudolf Otto in his (contested) classic *Das Heillige*, see the passage as emblematic of the Old Testament view that between the human and the divine spheres there is an ontological difference. Jacob does not attempt to ascend the ladder. On awakening he is seized with fear over having spent the night in a place of whose status he was unaware (p.8-9) [4]. This is the biblical plotline. In Grundtvig’s poem of the story, entitled ‘Jacob’s dream’ (1837) , however, the distance and thus the fear are notably underplayed. Jacob lies down “under open heaven” with a “stone as resting-pillow”. The moon is shining, and the stars are twinkling, as he lies down “in prayer to rest”; immediately he falls asleep “in heaven’s lodge”. In other words, he is at peace in the arms of nature (p.125-127) [19]. In his dream he sees “the King of all Kings”, who vows that Jacob will be the founder of “tribes throughout the earth”. When the vision is over, he exclaims, “that Heaven was so close I did not know!” He sets off “hopefully in the Lord’s name”, even while the lark is singing “sweetly in the morning sun”. He renames the place ‘Bethel’ and calls it “the Lord’s house in Heaven’s meadow”. This is a remarkably unproblematic, carefree image of a situation of contact between the godly and the human sphere. Notably one going against the grain of the tabu so clearly accentuated in the Old Testament version encouraging the intended reader to avoid the ‘glory of Jahve’ at any cost. For Grundtvig, however, contact between heavenly agents and humans are unproblematic, benign events to be sought after. In fact, he points to what he thought of as an infrastructure of communication between the different spheres: an everyday stairway to heaven is accessible through the “amazing musical scale” (tonestige) (Grundtvig 1825, 157) that hymnsinging represents. This is the center of Grundtvig’s understanding of the divine-human relationship which allows him to conclude that “the poetic element in humankind manifests itself in striving on Earth to a be reminded of Heaven” (p. 122) [14]. His hymns and songs seek to mediate the myriad earthly concerns with the divine presence behind them, as in the (for Skandinavians) well-known Christmas hymn, where the stairway of Jacob’s vision once again appears: “Lovely is the midnight sky,/beautiful to see on high,/where the golden stars are blinking,/where they smile in concert winking/us to join them up above.”

References

[1] R. Bartlett, Why Can the Dead Do Such Great Things?: Saints and Worshippers from the Martyrs to the Reformation, Princeton University Press, Princeton, NJ, 2013.

[2] J. F. Møller, Grundtvigs død Aarhus Universitetsforlag, Aarhus Universitetsforlag, 2019.

[3] J. Kjærgaard, Salmehåndbog, Det Kgl. Vajsenhus' Forlag, 2003.

[4] H. J. Lundager Jensen, Igennem urenhed til himlen. Den tanatologiske transformation fra israelitisk religion til kristendom, Religionsvidenskabeligt Tidsskrift 2019 (2019) 5–37.

[5] A. Korzybski, Science and Sanity: An Introduction to Non-Aristotelian Systems and General Semantics, 5th edition ed., Institute of General Semantics, Brooklyn, N.Y, 1933.

[6] E. Vylomova, L. Rimell, T. Cohn, T. Baldwin, Take and Took, Gaggle and Goose, Book and Read: Evaluating the Utility of Vector Differences for Lexical Relation Learning, 2016. URL: http://arxiv.org/abs/1509.01692. doi:10.48550/arXiv.1509.01692, arXiv:1509.01692 [cs].

[7] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, J. Dean, Distributed Representations of Words and Phrases and their Compositionality, 2013. URL: http://arxiv.org/abs/1310.4546. doi:10.48550/arXiv.1310.4546, arXiv:1310.4546 [cs, stat].

[8] X. Rong, word2vec Parameter Learning Explained, 2016. URL: http://arxiv.org/abs/1411.2738. doi:10.48550/arXiv.1411.2738, arXiv:1411.2738 [cs].

[9] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, E. Lefebvre, Fast unfolding of communities in large networks, Journal of Statistical Mechanics: Theory and Experiment 2008 (2008) P10008. URL: http://arxiv.org/abs/0803.0476. doi:10.1088/1742-5468/2008/10/P10008, arXiv:0803.0476 [cond-mat, physics:physics].

[10] R. N. Bellah, H. Joas (Eds.), The Axial Age and Its Consequences, 1st edition ed., Belknap Press: An Imprint of Harvard University Press, Cambridge, Mass, 2012.

[11] L. White, The Historical Roots of Our Ecologic Crisis, Science 155 (1967) 1203–1207. URL: https://www.jstor.org/stable/1720120, publisher: American Association for the Advancement of Science.

[12] A. Holm, To samtidige: Kierkegaards og Grundtvigs kritik af hinanden (ebog, pdf, dansk) af Anders Holm, Forlaget Anis, København, 2009. URL: https://www.williamdam.dk/to-samtidige-kierkegaards-og-grundtvigs-kritik-af-hinanden__85996.

[13] H. Wigh-Poulsen, Digteren og den sandheds ånd. Grundtvigs helligåndsteologi og den engelske romantik, Grundtvig-Studier 42 (1991) 68–85. URL: https://tidsskrift.dk/grs/article/view/16059. doi:10.7146/grs.v42i1.16059, number: 1.

[14] N. Grundtvig, [Anmeldelse af] Høimesse-Psalmer af B. S. Ingemann, Ph.D. thesis, 1825. URL: http://www.xn--grundtvigsvrker-7lb.dk/tekstvisning/16852/0#{%220%22:0,%22v0%22:0,%22k%22:0}.

[15] J. Høgenhaven, Grundtvig som fortolker af Det Gamle Testamente, Grundtvig-Studier 62 (2011) 51–80. URL: https://tidsskrift.dk/grs/article/view/16579. doi:10.7146/grs.v62i1.16579, number: 1.

[16] H. Grell, Skaberordet og billedordet. Studier over Grundtvigs teologi om ordet, Skrifter udgivet af Grundtvig-Selskabet XVII, 1980.

[17] D. C. McDannell, B. Lang, C. McDannell, Heaven: A History, Second edition, second edition ed., Yale University Press, New Haven, 2001.

[18] J. N. Bremmer, Christian Hell: From the Apocalypse of Peter to the Apocalypse of Paul, Numen 56 (2009) 298–325. URL: https://www.jstor.org/stable/27793794, publisher: Brill.

[19] I. L. Mikkelsen, Hyrdeliv og paradisdrøm. Om Grundtvigs syn på hyrder, Grundtvig-Studier 45 (1994) 122–141. URL: https://tidsskrift.dk/grs/article/view/16145. doi:10.7146/grs.v45i1.16145, number: 1.

Policy Issues vs. Documentation: Using BERTopic to Gain Insight in the Political Communication in Instagram Stories and Posts during the 2021 German Federal Election Campaign

Michael Achmann, Christian Wolff

Media Informatics Group, University of Regensburg, D-93040, Regensburg, Germany

Abstract

We give first insights in the political communication of the 2021 Federal election campaign in Germany. We focused on political messages found in ephemeral stories (n=2208) and permanent posts (n=718) shared in the last fortnight of the campaign. Topic modeling with BERTopic did not yield topics as finely grained as the categories of prior content analyses, yet two main themes emerged: The majority of posts deal with policy issues, while the majority of stories does not deal with policy issues. We found a large body of stories to be documentation of the rallies and campaign trail.

Keywords

topic modeling, social media analysis, instagram stories, visual social media, political communication

1. Introduction

Instagram, initially an image-sharing platform, is now one of the most popular social networks [1]. It has added features such as the algorithmic timeline, stories, and reels, all of which focus on visual media [2]. As such, it has become an important network to be used in election campaigns. The political communication of these campaigns, has been analyzed in several studies in the past [3]. The story feature, however, has attracted little attention from scholars even though the ephemeral character of stories stands out in a world of technology where "forgetting has become the exception, and remembering the default" [4]. In order to gain an initial understanding of an election campaign on Instagram, we analyzed the differences between stories and posts, proposing to focus on text-integrated images and classify the content using topic modeling. By comparing these two forms, we want to shed light on how political communication is evolving in response to changes in technology and social media usage and see our work as part of a larger body of research that seeks to understand the ways in which social media is changing political communication. Thus we try to answer the following questions:

DHNB2023 | Sustainability: Environment - Community - Data. The 7th Digital Humanities in the Nordic and Baltic Countries Conference. Oslo – Stavanger – Bergen, Norway. March 8–10, 2023.

michael.achmann@informatik.uni-regensburg.de (M. Achmann); christian.wolff@informatik.uni-regensburg.de (C. Wolff)
https://go.ur.de/michael-achmann/ (M. Achmann); https://go.ur.de/christian-wolff/ (C. Wolff)
0000-0002-4754-7842 (M. Achmann); 0000-0001-7278-8595 (C. Wolff)

DHNB Publications, DHNB2023 Conference Proceedings, https://journals.uio.no/dhnbpub/issue/view/875

1. What are the main political message types on the Instagram accounts of front-runners and political parties in the final two weeks of the 2021 election campaign?
2. How do these messages differ between ephemeral stories and permanent posts?
3. How well can we answer these questions computationally through the lens of BERTopic?

1.1. Political Communication on Instagram

Despite the young age of Instagram, the political communication on the platform has already been studied in numerous papers, with a focus on different political actors and different nations: Bast reviewed 37 studies, systematizing them according to the methodological approaches, theories, and sampled data. She found studies to address three key areas: "Who uses Instagram, how do they use it, and with what effect?" [3]. The majority of studies employed a quantitative approach, with quantitative content analysis being the most prevalent methodology utilized. The variety of approaches and theories lead to a multitude of study designs: Some studies compare communication strategies on different platforms (e.g. [5]) or between party accounts and politician accounts (e.g. [6]). Others, such as Lalancette and Raynauld's highly-cited analysis of Justin Trudeau's Instagram use, which utilizes the theoretical framework of celebrity politics, concentrate on a single individual [7].

The overarching result of these studies shows political figures are taking advantage of Instagram to present a positive and encouraging image, rather than delving into policy issues, meeting with constituents, or organizing voting efforts. Most posts feature pictures of the political actor themselves or, in the case of political party accounts, images of their leading candidate. A comparison to a similar literature analysis focusing on Twitter usage during political campaigns reveals similar usage styles: both platforms are rarely used by politicians to interact with voters, though there is considerable variance between individual players.

Since the literature review's publication several new studies about political communication using Instagram have been published, for instance a first longitudinal study using *CrowdTangle*[1] to retrospectively collect posts to shed light on changes in Instagram use of European political parties over time [8]. Further, studies regarding the visual communication of European right-wing populist politicians [9], differences in user engagement with political parties between Instagram, Facebook and Twitter in Canada [10], Instagram use of Spain's major political parties [11], a cross-country study of politicians' self-depiction [12] using computer vision, once more Justin Trudeau's use of Instagram [13] and finally the use of Instagram stories by Trump and Biden in the 2020 presidential election [14] have recently been published.

Bast concludes her review study arguing in favor of more systematic comparisons with larger and more heterogeneous samples, as well as more longitudinal studies that go beyond single election campaigns. She suggests that the lack of precise and coherent definitions of concepts and content analyzed be remedied by transferring established analytical concepts in order to build solid evidence. In addition, she argues that the relatively new Instagram functions *Instagram Video* and *Instagram Stories* present a valuable opportunity for further research, an argument backed by others [6, 14]. Stories are distinct from posts, which are the original content

[1] https://www.crowdtangle.com/

on Instagram. Posts can include one or more images and / or videos that are permanently shared on a user's profile, often accompanied by captions (text content). In contrast, stories are a relatively new feature (see below) that are ephemeral and solely composed of an image or video. Unlike posts, they disappear after a certain period of time and do not remain on a user's profile.

1.2. Computational Analysis of Social Media Content

We see potential to increase the comparability of social media analyses through the use of computational methods to create reproducible and valid analyses. In addition, computational approaches enable us to handle a growing amount of user generated content [15, ch. 1], namely visual content in the context of Instagram. We propose to focus on text-integrated images and captions in order to apply computational text analysis methods, which are well established [16], and may serve as a bridge towards the computational analysis of visual media, which is yet a challenge [17, 18]. Overall, we want to explore the potential of computational approaches to discover and analyze visual social media content, with present work focusing on topic modeling as one possible candidate in the development of a workflow for computational visual content analysis.

1.3. Topic Modeling & Instagram

While Instagram is primarily focused on visual media, text has already been used to explore themes of posts: Rodina and Dligach analyzed the themes and topics of posts by Ramzan Kadyrov, dictatorial head of the autonomous Chechen Republic. Using the Latent Dirichlet allocation (LDA) model they found two dominant themes across 6854 analyzed posts and 24 topics: A personal and a political theme, which over time started blending [19]. In health domains several social media studies relied on topic modeling: Murashka et al. tried to identify objectification elements from image captions and comments of popular #fitspiration accounts. Kim et al. [21] identified how people managed their daily lives in the face of the pandemic's fear and discomfort by applying topic modeling on captions and image descriptions. Similarly Muralidhara and Paul [22] looked into Instagram posts with health-related hashtags and identified 47 health-related topics in their corpus. They trained their model on hashtags and caption words to automatically generate image tags. In a journalistic context Al-Rawi et al. [23] employed topic modeling in a mixed-methods approach to explore the most liked news topics across several news accounts.

1.4. Ephemeral Content & Stories

While a research gap in political communication exists, ephemeral Instagram stories have been investigated in other disciplines. Stories are a special type of post as they expire after 24 hours and became the platform's main growth engine [2]. After expiry, they are archived for the authoring user but not for other users. They consist of videos or images, or collages of media and so-called stickers, platform specific affordances [24] to tag other users; hashtags; locations; or allow for interaction through e.g. questions or quizzes. Since Snapchat invented the ephemeral feature, it is worth to look at Rettberg's study of Snapchat content. She suggests that the app changed online communication and its affordances enable the discovery of more conversational methods of communicating and telling stories, thus "Snapchat is a conversation, not an archive". Through qualitative content analysis, observation and in-depth interviews, Amancio found four narrative elements used by Snapchat and Instagram storytellers to tell their stories and construct a narrative: actions (demonstrating emotions, eating, interacting), happenings (updates), characters (people, self-portraits and animals) and setting (environment), making use of images, texts, videos, emoji, doodles, instant information and filters [26]. Bainotti et al. investigated 292 Instagram Stories by private users using an ethnographic coding approach. They claim to have identified specific grammars by matching the content and context-of-use, the two main ones are: "a grammar for documentation and a grammar for interaction" [27]. Other areas of interest for stories were ephemeral journalism [28] and Female Atheletes' self-presentation [29]. Closer to political communication, a study of the candidates for the 2016 U.S. presidential primaries identified ten frames used on Snapchat [30]. Finally, Towner and Muñoz [14] published a first analysis of political communication in Instagram Stories, studying the stories published by the two U.S. presidential candidates in the 2020 campaign. They collected a sample of 304 images one week before and after the election campaign. From a marketing perspective, they saw several flaws, like missed opportunities of sharing user-generated content and inconsistently following communication norms for Instagram Stories. Further, campaign events and rallies were the most popular type of messages.

2. Methods

In order to uncover the main political messages of posts and stories in the 2021 election campaign, we used word frequency measures, word clouds and topic modeling, an unsupervised machine learning technique. Since Instagram primarily consists of visual media, we applied optical character recognition (OCR) to translate text-integrated images into machine readable text.

2.1. Data Collection

We collected a sample of 2208 stories and 718 posts shared by politicians and parties within the last fortnight of the 2021 federal election campaign. Stories were collected daily at 0:00 (CET) using `Selenium`, a Python package to simulate a human user browsing the stories.[2] Posts were collected retrospectively through *CrowdTangle* and `Instaloader`. Germany's multiparty system has witnessed a growing trend of fragmentation in recent years. As a result, we conducted a comprehensive data collection, focusing on posts and stories shared by the eight political parties participating in the election, which currently hold seats in state or federal legislatures and possess verified Instagram accounts (refer to Table 1). Additionally, we ensured the inclusion of at least one front-runner from each party (refer to Table 2).

2.2. Preprocessing

There are three different sources for text which we we have used in our analysis: 1) post captions, which is computer readable text added by users to posts, 2) text-integrated posts, and 3) text-

[2] Data for Sep 14 is incomplete due to technical problems. For present proof of concept work the incompleteness of the sample has been ignored.

Table 1
Selected parties and their Instagram handles at the time of data collection.

Party (Abb.) @handle	Party (Name) Translation
AfD @afd_bund	Alternative für Deutschland Right-wing Populist Party (Alternative for Germany)
CDU @cdu	Christlich Demokratische Union Deutschlands Centre-right, Christian Democrats (Christian Democratic Union of Germany)
CSU @christlichsozialeunion	Christlich Soziale Union in Bayern Bavarian Centre-right (Christian Social Union)
Die Grünen @die_gruenen	Bündnis90 /Die Grünen Green, Environmental Politics (Alliance 90/The Greens)
Die Linke @dielinke	Die Linke Democratic Socialists, Left-wing (The Left)
FDP @fdp	Freie Demokratische Partei Classical Liberals, Pro-business Free Democrats (Free Democratic Party)
FW @fw_bayern	Freie Wähler Centrist, Citizens' Groups (Free Voters)
SPD @spdde	Sozialdemokratische Partei Deutschland Social Democrats, Centre-left (Social Democratic Party of Germany)

integrated stories. The text of the latter two is embedded in either images or videos. While captions and embedded text for posts are available through *CrowdTangle*, the embedded text in stories is not. Thus we used EasyOCR for Optical Character Recognition – for consistency on both, stories and posts – to extract the embedded text from text-integrated images. A majority of stories (n=1246) turned out to be videos. As the Instagram app just allows to add one combination of text and stickers which is displayed across all frame of videos, we extracted the first video frame using OpenCV.[3]

2.3. OCR & Relevance Classification

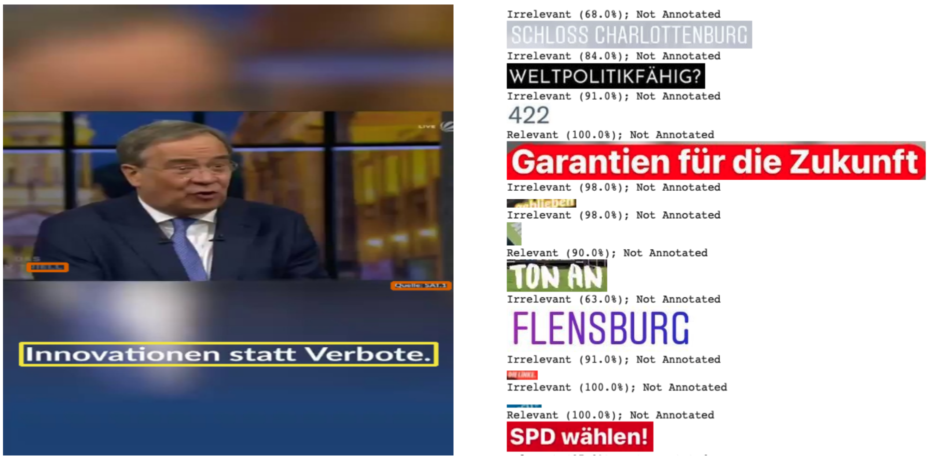
Since EasyOCR turned out to be overambitious recognizing the embedded text, e.g. transcribing shop signs from the image's backdrop, we trained a CNN[4] to classify relevant and irrelevant text-snippets (see figure 1). A human annotator corrected the OCR results and annotated the relevance of each text snippet for 50% of all captured stories. Sticker content has been labeled as irrelevant since their content is available in the metadata. Through the annotation process more than half of the OCR annotations were deemed irrelevant (4794 out of 9850). Our model reached an f1-score of .94 which we deemed sufficient (see table 3). At the same time, only

[3] This approach disregards embedded text in videos itself, like subtitle. We see future work taking every frame into account, controlling for repeated text across frames.

[4] A Convolutional Neural Network, a type of neural network used in machine learning to classify images. After some experiments we archived the best results using only the cropped images of text snippets as input data.

Table 2
Selected politicians' accounts and their positions and party affiliation at the time of data collection. (The party GRÜNE is referenced as B90DieGruenen later on.)

Name	Party	Position	@handle
Alice Weidel	AfD	Front-Runner	@alice.weidel
Jörg Meuthen	AfD	Head of Party	@joerg.meuthen
Armin Laschet	CDU	Chancellor Candidate	@armin_laschet
Markus Söder	CSU	Head of Party	@markus.soeder
Annalena Baerbock	GRÜNE	Chancellor Candidate	@abaerbock
Robert Habeck	GRÜNE	Front-Runner	@robert.habeck
Ates Gürpinar	Die Linke	Deputy Head of Party	@atesgurpinar
Susanne Henning-Wellsow	Die Linke	Head of Party	@susanne_hennig_wellsow
Christian Lindner	FDP	Front-Runner	@christianlindner
Nicola Beer	FDP	Deputy Head of Party	@nicola_beer
Engin Eroglu	FW	Deputy Head of Party	@engin_eroglu
Gregor Voht	FW	Deputy Head of Party	@grey_gor
Olaf Scholz	SPD	Chancellor Candidate	@olafscholz
Saskia Esken	SPD	Head of Party	@saskiaesken



```
Irrelevant (68.0%); Not Annotated
SCHLOSS CHARLOTTENBURG
Irrelevant (84.0%); Not Annotated
WELTPOLITIKFÄHIG?
Irrelevant (91.0%); Not Annotated
422
Relevant (100.0%); Not Annotated
Garantien für die Zukunft
Irrelevant (98.0%); Not Annotated
Irrelevant (98.0%); Not Annotated
Relevant (90.0%); Not Annotated
TON AN
Irrelevant (63.0%); Not Annotated
FLENSBURG
Irrelevant (91.0%); Not Annotated
Irrelevant (100.0%); Not Annotated
Relevant (100.0%); Not Annotated
SPD wählen!
```

Figure 1: Application of the trained model for relevance classification. The rescaled image of a story on the left shows yellow bounding boxes for relevant and red ones for irrelevant text snippets. The right-hand image shows a step of the processing pipeline: Classification takes place for each extracted snippet. The top line, for example, shows the correct classification of a location sticker as irrelevant since we are able to extract the sticker's content from the metadata.

minor human adjustments were required for the OCR of relevant text-snippets: Using R's `adist` (approximate string distances), we calculated the generalized Levenshtein-Distance between OCR and human-corrected text: there was a mean distance of only .51 characters. Overall 104 transcriptions (across 57 pictures) were added entirely from scratch throughout the annotation process.

Table 3
Classification report of the classification model.

	precision	recall	f1-score	support
Irrelevant	0.97	0.93	0.95	5537
Relevant	0.91	0.96	0.94	4349
accuracy			0.94	9886
macro avg	0.94	0.94	0.94	9886
weigthed avg	0.94	0.94	0.94	9886

All text has been preprocessed using the `text-clean` python package to tackle encoding discrepancies and to remove emojis. For the word clouds, German characters have been converted to ASCII characters, punctuation has been removed and finally stop words were deleted using NLTK's German stopword list.

2.4. Topic Modeling

In order to gain a computational insight of our corpora, we employ topic modeling. It is a set of algorithms that assist in identifying recurring themes in a corpus of documents [31]. Probabilistic models like the Latent Dirichlet Allocation (LDA) assume that each document exhibits multiple topics in different proportions [32]. A topic is a set of terms which frequently co-occur across the documents. These topics and corresponding terms aid in uncovering themes across the set of documents, in the case of Instagram posts and stories we employ topic modeling to help uncovering the content and identify policy issues of the 2021 election's Instagram campaign. The result of our topic model(s) serves as the basis for the message type analysis.

Considering recent developments in language models, we sought out approaches based on modern language models. We identified *BERTopic* [33] as suitable software, since the author reports BERTopic to perform well on a corpus of tweets. Egger and Yu compared several topic modeling approaches and confirm the performance on tweets. While we are dealing with Instagram content, tweets are by definition social media content and contain short texts, thus they are comparable to our corpora.

Once the relevance model was applied to the two OCR corpora of text-integrated posts and stories, we created an overarching corpus consisting of the captions and relevant image-text. Multiple text lines of image text were concatenated to one document, thus one story image corresponds with one document in the corpus and one post corresponds to at least two documents, one for the caption, and one or more for the image text (with one post corresponding to one or more images). We trained the following three models:

Overall Model The overall model is our naive starting point. It was trained using all documents from the corpora together, thus it includes post and story OCR as well as post captions. We used the `fasttext` [35] German word vector model `cc.de.300.bin` and `paraphrase-multilingual-MiniLM-L12-v2` sentence transformer. As per BERTopic documentation, we removed stop words using `scikit-learn`'s `CountVectorizer` method. NLTK's German stop word list was applied after the word embeddings have been created and documents were clustered.

Post Model Once we spotted first weaknesses in the overall model (see below), we decided to train models per corpus. First tests using either OCR or caption documents yielded a low amount of topics. Consulting the BERTopic documentation, we decided to split captions by sentences to generate a larger set of documents. As hashtags and mentions started dominating the topics, we added an additional preprocessing step, removing any hashtags and mentions from the caption corpus.

Story Model The story model was trained using the same parameters as the overall model. It was only trained using the story OCR documents.

2.5. Political Messages

In a second step we use the results of the topic modeling as a basis for message type classification. Due to the topic modeling results we suggest to categorize the messages into two main categories along the existence or absence of policy issues. Towner and Muñoz introduce ten main types with higher granularity: campaign events, thank you, voter endorsement, policy issues, hybrid, character, attack, behind the scenes, mobilization and other [14]. They based their categories on Liebhart and Bernhardt's study of an Austrian presidential campaign [36], which also served as a foundation for Haßler et al.'s [6] image types. The later developed nine image types: policy, campaign events, call for action, negative campaigning, media work, campaign material, supporters, everyday political work, private background story. While our results of the unsupervised topic modeling approach did not yield topics as finely detailed as to sharply distinguish between the messages types from the literature, we are optimistic that the two types will be sufficient to elucidate the differences between stories and posts.

Policy Issues We considered policy issues to be actual political content regarding a variety of domains. A post or story was considered to regard policy issues if any issue was present at all. Fringe cases (e.g. short sentences by B90DieGruenen using their neologism "Klimaregierung" (climate government), calling for more climate action) were considered as policy-bearing while slogans (e.g. "Entlasten statt Belasten", relieve rather than burden, by CSU) by itself was not considered to refer to policy issues, since the first example explicitly takes up a political issue, while the slogan in the second example merely refers to abstract relief without any specific indication of what type of relief is meant.

No Policy Issues This category is the opposite of the policy issue category and merely describes the absence of political topics in the textual content. Examples include text-integrated posts bearing "Damit es möglich wird: Am 26.9. DIE LINKE wählen" (to make it possible: vote for DIE LINKE on 26.9.) or "Danke für eure Unterstützung" (thank you for your support) posts. It includes the types campaign events, thank you, voter endorsement, character, attack, behind the scenes, mobilization and other by Towner and Muñoz.

3. Results

Throughout the period of investigation, a total of 2208 stories were collected, most of them (n=1246) videos. In the same period 713 posts were published by the parties and politicians. While B90DieGruenen created the most stories (n=578), the CDU published the most posts (n=144). The most stories were created on Sept. 24 (n=311), and the most posts on Sept. 23 and 24 (n=75 on both days), with an average of 157.71 stories and 50.93 posts published per day. The application of the preprocessing steps described above resulted in three text corpora derived from the Instagram content: 1. captions, 2. post OCR, and 3. story OCR.

Captions are genuine digital text added by the user when creating a permanent image or video post. 706 posts were accompanied by captions (99%). Unsurprisingly, captions were the longest texts we observed with an average of 75.80 (median=60.00) words per post. Each caption incorporated an average of 2.72 hashtags (median=2.00). Among the ten most frequent words across all posts were terms like "deutschland" (Germany), "heute" (today), "wählen" (to elect), "stimmen" (to vote), "bundestagswahl" (federal election). Before filtering the hashtags "btw21', "csu" and "cdu" ranked among the most top words, hinting at a consequent use of these hashtags by the CSU and CDU parties in combination with them being the most active posters.

Post OCR consists of the OCR results classified as relevant. Each post may contain several images or videos[5]. We collected a total of 1299 image files belonging to 713 posts. `easyocr` identified text in 1092 images, out of which 650 images contained text classified as relevant.[6] Overall, the 650 images belong to 532 different posts, thus 74.61 % of posts contain text-integrated images. On average, text-integrated post images contain 15.87 words (median=12.00). The most frequent words look similar to the captions with "deutschland" (Germany), "wählen" (to elect), "stimmen" (to vote) taking the top places.

Story OCR comprises the relevant text snippets found in stories, with each story containing one image that may include multiple text snippets. Further, stories offer the possibility to add so-called stickers, like hashtags, mentions or locations. Our relevance classification model was trained to reject text from the stickers since we can retrieve sticker information from the metadata. Out of a total of 2208 stories 1939 (87.82%) contain relevant text-snippets. On average each story contains 15.66 (median=12.00) words. The most frequent term is "heute" (today), followed by "uhr" (o'clock) and "danke" (thank you). While "deutschland" (Germany) and "wählen" (to elect) also rank among the top ten words for stories, their relative frequency is lower in comparison to captions and post OCR.

Overall, we see text to be part of the majority of shared content. Almost every post contained a caption and the majority of post images and stories were text-integrated. Top word frequencies show first similarities and differences across the corpora. In order to gain a better insight of the

[5] We used the cover image of videos for our analysis.

[6] Since we trained our model using annotated story images we qualitatively explored the rejected text snippets and are confident not to have missed important content. In addition, we took a deeper look at images without any text (after relevance classification) and did not find a single text-integrated image rejected erroneously.

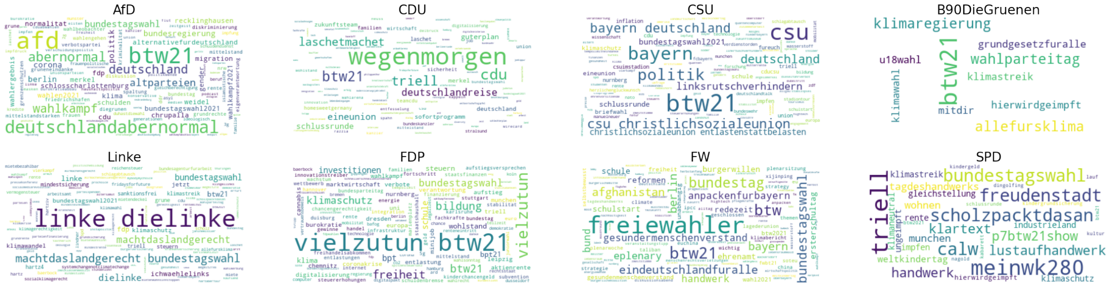


Figure 2: Word cloud of hashtags used in captions grouped by parties.

content shared by politicians and parties we first take a look at the word clouds and thereafter present an overview of topics as generated by BERTopic.

3.1. Looking at the Word Clouds

One platform-specific affordance [24] of Instagram is the use of hashtags. While hashtags may be used in both, stories and posts, our data shows that hashtags are rarely used in stories. However, since they have been used in the past [37] for automatic image annotation, we will first take a look at caption hashtags through word clouds and then proceed to our three corpora. The word clouds have been generated using the `WordCloud` package. We grouped the text documents by party, this split allows, on the one hand, to see differences between the parties, and, on the other hand, to control the dominance of parties with a higher posting frequency.

Hashtags Most parties referenced policy issues in their hashtags (e.g., “#klimaschutz” (climate protection), “#digitalisierung” (digitization), “#impfen” (vaccination). FDP used the most varied hashtags, followed by SPD and B90DieGruenen. Across all parties the #btw21 was the most used hashtag. Non-policy related hashtags, including party names and slogans (e.g., “#wegenmorgen” (abouttomorrow), “#vielzutun” (lotstodo), “#deutschlandabernormal” (germanybutnormal)), were dominant for all parties. The CSU appears to be the only party to have established a negative-campaigning hashtag “#linksrutschverhindern” (see figure 2).

Captions Similar to the hashtags, caption word clouds rarely offer insight into policy issues important for the campaign. Mentions of election, Germany, party names and front-runners were frequent. CSU was the only party with more occurrences of “bayern” (Bavaria). B90DieGruenen most consistentently referenced their posts with “klimaschutz” (climate protection). Other policy issues mentioned were “hartz IV” (social welfare benefit program, Linke), “familie” (family, CSU) and “schulen” (schools, FDP).

Post OCR The text-integrated posts focus on party names and the election, with the AfD associated with "nein" (no) possibly indicating negative campaigning. FDP refers to "bildung" (education), Linke to "pflege" (nursing), "rente" (pensions) and "klimaschutz" (climate protection). CSU’s slogan is "entlasten statt belasten" (better relieve than burden) focusing on Bavaria alongside FW.

Story OCR The term “heute” (today) dominates stories across parties, as do front-runner names. The AfD used a question sticker, resulting in that question dominating their stories. The FDP held a political convention which was referenced in many stories.

The macroscopic look at the wordclouds allowed us to gain a first insight into the Instagram election campaign. Overall, we see some policy issues emerging. Mostly, however, they are obfuscated by the parties mentioning their names, referencing their candidates and using slogans or tailored hashtags again and again, thereby dominating the word frequencies which are the basis of word clouds. Thus, in order to computationally gain a better insight of the campaign we used BERTopic for topic modeling.

3.2. Looking through the lens of BERTopic

Once the first model has been trained, several policy-issues appeared among the topic representations. Over the course of several iterations we refined the model(s) and decided to split posts and stories in order to gain better results. Once we trained the overall model we inspected the topics through a dendrogram and an inter-topic distance map and decided to reduce the initial topic count of 62 to 25. Topic -1 refers to outliers, all items sorted into this topic were disregarded. For comparability we kept the topic count constant across the three models.

First, we offer a qualitative look into the results for each topic model. In a second step during qualitative inspection we assigned a new variable to each topic, in order to distinguish topics dealing with policy issues from topics with other, non-policy related issues. While a majority of topics clearly leaned towards the absence or presence of policy issues, some could not be subsumed in either of the classes. These were classified as mixed, in order to prevent misinterpretations. Since each post consists of one caption and at least one image, we considered a post to contain policy issues if at least one image or the caption contained policy issues. Thereafter we are going to use this variable for a quantitative interpretation of our corpora.

Overall Model The overall model uncovered several policy issues, like climate change, renewable energies, education & digitization, labour and social issues, and the economy. On a closer look, however, several topics showed inconsistencies between the different corpora. One topic, for example, consists of text-integrated posts with short texts about policy issues, as well as a majority of policy-focused captions. The stories in this topic, however, are mostly documentation. Further, several policy issues were mixed together in plenty of topics, thus a differentiation of policy issues by the overall topic model may not be very accurate. All in all on visual inspection of the post images and stories we saw first patterns emerge, namely the difference between policy-issues focused content and documentary content. Further, the topic representations appeared in several cases to only cluster either posts (OCR / caption) or stories into meaningful categories, we call these *mixed* topics. Hence we decided to train separate models to gain better insight, hoping to improve topic validity.

Stories Model The story model uncovered several clusters of stories documenting the election campaign, namely some "thank you!" and "hello" topics as well as a city names / geographical locations topic and a "selfie & beer garden" topic. One topic identified mostly

Table 4
Overview of the 25 topics in the posts model, the amounts of post captions per topic and the message type variable assigned for each topic.

Topic	Post Captions		Title	Message Type
	Party	Person		
0	47	28	Climate, CO2 and more	Pol. Issues
1	32	10	Focus on Germany, mostly Slogans	Mixed
2	2	2	Announcement and Angela Merkel	No Pol. Issues
3	6	2	Election-Centred	No Pol. Issues
4	1	2	Need to Fight	Mixed
5	33	5	CDU / CSU Slogans, Mixed with Policy Issues	Mixed
6	13	1	Election Day	No Pol. Issues
7	0	0	NA	No Pol. Issues
8	5	2	Thanks, Thanks, Thanks	No Pol. Issues
9	18	5	CSU Slogan, Financial Relief	Pol. Issues
10	0	0	NA	No Pol. Issues
11	13	12	Democracy, Middle Class, and more	Pol. Issues
12	0	3	Announcements	No Pol. Issues
13	11	4	Focus on the Union (CDU+CSU Party)	No Pol. Issues
14	4	0	Crisis and Scandals, mixed with negative-campaigning	Mixed
15	2	0	FDP-Slogan	Mixed
16	34	16	Digitization and Education	Pol. Issues
17	90	60	Financial Relief, Taxes, Debt: Share Pics with short Policy Snippets, mixed with slogans	Mixed
18	5	1	Specific Amount of Money	Pol. Issues
19	3	1	Children and Families	Pol. Issues
20	19	11	Schools and Education	Mixed
21	4	11	COVID and Reliability	Pol. Issues
22	96	72	A variety of political issues, centered around the change for the future	Pol. Issues
23	6	7	Change and Future Direction	Pol. Issues
24	5	9	The gap beetween Cities and Countryside, Germany-Centred	Pol. Issues
Total	449	264	713	

negative campaigning and another one stories focused on the chancellor candidates. While a majority of stories appeared in these documentation topics, we also discovered several hybrid topics, which consist of stories documenting campaign rallies or similar events supplemented by quotes or short snippets referring to policy issues. Namely taxes, economy, progress, social, education and children appeared as policy issues among these topics. They are different to the *mixed* topics, as they are not a mix of stories with and without policy issue, rather they combine the typical elements of documentation style images with policy issue through e.g. through short quotes, thus we assessed these stories as policy issue message types. Finally, a few policy-focused topics emerged: We observed climate change, democracy, young people, family and children as policy issues through the lens of this model.

Posts Model The initial, unreduced, posts-model based on caption-sentences offers the most detailed look into policy issues: climate change, digitization, education and family emerge, moreover we see issues like employment, affordable housing, COVID-19, police and safety, debt and economical directions, democracy, creative and cultural industry, deregulation and the reduction of bureaucracy, and agriculture and the countryside. Once the model was reduced to the target of 25 topics, some themes such as Climate change, democracy, education, schools and digitization, and children and families remained clearly visible .In

Table 5
Overview of the 25 topics in the story model, the amounts of stores per topic and the message type variable assigned for each topic.

Topic	Stories	Title	Message Type
0	107	Thanks, Thanks, Thanks	No Pol. Issues
1	135	Climate Change	Pol. Issues
2	23	Documentation	No Pol. Issues
3	72	Election & Candidates	No Pol. Issues
4	73	Documentation	No Pol. Issues
5	202	Place Name & Documentation	No Pol. Issues
6	32	Democracy, Young People, Family, Hatred	Pol. Issues
7	108	Mostly Documentation	Mixed
8	222	Short Policy Issues (Taxes, Economy) and Documentation	Pol. Issues
9	37	Pol. Issues of AfD and Die Linke; others: Documentation	Mixed
10	27	Interviews	No Pol. Issues
11	9	Party Names, Negative Campaigning	No Pol. Issues
12	38	Selfies & Beer Garden: Documentation	No Pol. Issues
13	16	Children & Family	Pol. Issues
14	47	Focus on Chancellor Candidate	No Pol. Issues
15	12	Documentation	No Pol. Issues
16	1	NA	No Pol. Issues
17	98	Documentation & Short Policy Snippets	Mixed
18	21	Progress & Persuasion, Documentary Hybrid	Pol. Issues
19	18	Interviews	No Pol. Issues
20	17	Hello, Hello, Hello!	No Pol. Issues
21	115	Press Conferences & Announcements, paired with Content	Mixed
22	42	Thanks, Thanks, Thanks	No Pol. Issues
23	192	Documentation	No Pol. Issues
24	216	Documentation, Minority Issues: Social, Education, Children	No Pol. Issues
Total	1880		

the end, the majority of posts are dealing with policy issues. However, seven topics had to be classified as **mixed** since they neither showed a clear majority of policy issues nor documentation-style images and captions.

Through the qualitative inspection of the three models and their topics we were able to gain an insight into several policy issues occurring through the posts and stories. The topics, however, did not always differentiate the posts / stories precisely into a coherent theme. Captions classified by the posts model, for example, are tough to group into one topic as they often contain one or more policy issues. At the same time the stories may be grouped rather well into a coherent topic due to the rather short text-length and focus on one issues, yet the majority of stories turned out not to deal with any policy issue at all.

All in all, we found a majority of posts to deal with policy issues (see figure 3, 64.24%). Almost

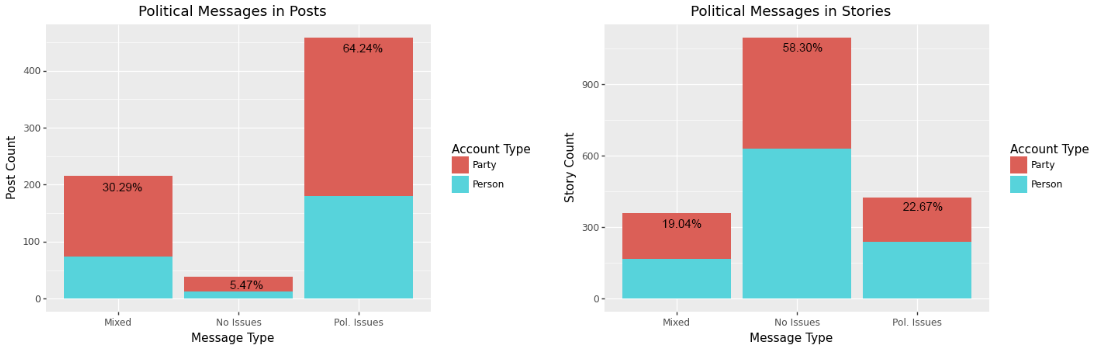


Figure 3: Policy issues in posts (left, n=713) against policy issues in stories (right, n=1880) by account type.

a third of posts (30.29%), however, fell into mixed topics, which could not be clearly classified as content bearing or not. Nevertheless, a clear minority of posts was identified to not deal with policy issues at all (5.47%). The stories, on the other hand, show a majority of items to not deal with (58.30%) policy issues. A smaller share clearly were regarded as policy-bearing content (22.67%) and a minority as mixed (19.04%). We have not observed any significant differences between user types (party accounts vs. political leaders), nor between different parties.

This story model unearthed what could be described as a subtype of both main message types: a documentation class. Exploring the stories of these documentary topics we find parallels to the findings by Bainotti et al. [27] about personal Instagram stories. They discovered a "grammar for documentation" that is used to portray both exceptional event photographs and regular, everyday situations. In the documentation topics, stories about political rallies showed campaign events, different stops along the campaign trail, front-runners on stage, and the crowd or both without any policy issues in the textual content. As such it consists of the types campaign events, and thank you in combination with campaign events by Towner and Muñoz. Nevertheless, there are some documentation topics in which a documentary style stories mix with short quotes or text snippets bearing references to policy issues. While we discovered the documentary themes almost exclusively among text-integrated stories, we took a step back and investigated the post images without text, which had eluded the attention of our text-based approach. The vast majority of those images appear to be documentary-style images without any policy issues.

4. Discussion

We presented a pipeline to explore Instagram posts and stories using tools for textual analysis. By analyzing word frequencies, constructing word clouds, and employing topic models we were able to gain an initial understanding of the 2021 German federal election campaign. We found that posts tended to focus on policy issues while stories typically did not. Further, we found a large share of stories to document the campaign trail of candidates and campaign events, consistent with a previous study of the U.S. presidential campaign. The proposed pipeline provides a valuable tool for exploring the rapidly increasing amounts of visual social media content posted during campaigns, despite topic modeling not yielding precise topics to differentiate between policy issues or messages types as finely grained as in previous work: For example, Haßler et al. collected 581 Instagram posts from the last month of the 2017 German federal election, while we have already gathered 713 posts comprising 1299 images and 2208 stories over the last fortnight of the 2021 campaign. Thus, while our results still call for a proper validation, e.g. through (computational) content analysis, we are able to give valuable first insights into stories used in a German election campaign. Consistent with Towner and Muñoz's [14] findings about the 2020 U.S. presidential campaign, we found stories to mostly consist of the no policy issues types, several topics discovered consist of stories documenting the campaign, similar to campaign events and rallies being the most popular message type in the U.S. elections.

The majority of past analyses of the political communication on Instagram relied on content analysis taking into account both modalities, images and text, in some cases even video and audio of the Instagram content. Our approach concentrated on text, and overall BERTopic turned out to be a capable aid in exploring Instagram content, since the share of text-integrated images was overall high. Thus, we were able to use this text-based approach for the majority of the content available. Our approach may be applied in different domains as long as there is a large amount of text-integrated images. Through our experiments, however, one shortcoming became apparent: Parties and politicians used several catch-phrases and slogans over and over again. These eventually started to appear as their own topics, thus some policy issues, especially the ones of parties (like the CDU) extensively using the same or similar phrasing over and over, did not develop as their own topics.

4.1. Limitations & Future Work

One limitation of our research is that we have only focused on images, when in fact more than half of the stories were videos. This could potentially lead us to miss important content that may be part of the audio channel or text embedded in subsequent video frames. In order to address this issue, future studies should consider using automated transcriptions in order to access another layer of textual content which could then be integrated into the proposed pipeline. Similarly, we focus not on the image itself, but on the text as part of it. This means our analysis overlooks important visual cues and content present in the picture. We could image the use of automated image descriptions (e.g., produced by CLIP [38]) in conjunction with captions and image text to bridge the gap between our topic modeling results and the granularity of quantitative content analysis seen in prior work. A similar approach has already been demonstrated to be effective by Muralidhara and Paul using automated image tags. Further, we see potential in using BERTopic's guided topic modeling capabilities in order to fine-tune the topics and allow for a more distinct separation between topics in order to capture policy issues and message types with higher validity. Such an approach has already been successfully used to study political topics in tweets [39]. Such a guided topic model could prove to be an invaluable asset for longitudinal studies of political communication on Instagram, a research desideratum that has been formulated in the literature.

References

[1] Social Networks nach Nutzern 2022, ???? URL: https://de.statista.com/statistik/daten/studie/181086/umfrage/die-weltweit-groessten-social-networks-nach-anzahl-der-user/.

[2] T. Leaver, T. Highfield, C. Abidin, Instagram: Visual Social Media Cultures, John Wiley & Sons, 2020.

[3] J. Bast, Politicians, Parties, and Government Representatives on Instagram: A Review of Research Approaches, Usage Patterns, and Effects, Review of Communication Research 9 (2021). URL: https://www.rcommunicationr.org/index.php/rcr/article/view/108.

[4] V. Mayer-Schönberger, Delete: The Virtue of Forgetting in the Digital Age, Princeton University Press, 2011.

[5] X. Farkas, M. Bene, Images, Politicians, and Social Media: Patterns and Effects of Politicians' Image-Based Political Communication Strategies on Social Media, The International Journal of Press/Politics 26 (2021) 119–142. URL: https://doi.org/10.1177/1940161220959553. doi:10.1177/1940161220959553.

[6] J. Haßler, A. S. Kümpel, J. Keller, Instagram and political campaigning in the 2017 German federal election. A quantitative content analysis of German top politicians' and parliamentary parties' posts, Information, Communication and Society (2021) 1–21. URL: https://doi.org/10.1080/1369118X.2021.1954974. doi:10.1080/1369118X.2021.1954974.

[7] M. Lalancette, V. Raynauld, The Power of Political Image: Justin Trudeau, Instagram, and Celebrity Politics, The American behavioral scientist 63 (2017) 888–924. URL: https://doi.org/10.1177/0002764217744838. doi:10.1177/0002764217744838.

[8] A. Olof Larsson, The rise of Instagram as a tool for political communication: A longitudinal study of European political parties and their followers, New Media & Society (2021) 14614448211034158. URL: https://doi.org/10.1177/14614448211034158. doi:10.1177/14614448211034158.

[9] J. Bast, Managing the Image. The Visual Communication Strategy of European Right-Wing Populist Politicians on Instagram, Journal of Political Marketing (2021) 1–30. URL: https://doi.org/10.1080/15377857.2021.1892901. doi:10.1080/15377857.2021.1892901.

[10] S. Boulianne, A. O. Larsson, Engagement with candidate posts on Twitter, Instagram, and Facebook during the 2019 election, New Media & Society (2021) 14614448211009504. URL: https://doi.org/10.1177/14614448211009504. doi:10.1177/14614448211009504.

[11] A. Pineda, E. Bellido-Pérez, A. I. Barragán-Romero, "Backstage moments during the campaign": The interactive use of Instagram by Spanish political leaders, New Media & Society 24 (2022) 1133–1160. URL: https://doi.org/10.1177/1461444820972390. doi:10.1177/1461444820972390.

[12] M. Haim, M. Jungblut, Politicians' Self-depiction and Their News Portrayal: Evidence from 28 Countries Using Visual Computational Analysis, Political Communication 38 (2021) 55–74. URL: https://doi.org/10.1080/10584609.2020.1753869. doi:10.1080/10584609.2020.1753869.

[13] V. Raynauld, M. Lalancette, Pictures, filters, and politics: Instagram's role in political image making and storytelling in Canada, Visual communication quarterly 28 (2021) 212–226. URL: https://www.tandfonline.com/doi/full/10.1080/15551393.2021.1986827. doi:10.1080/15551393.2021.1986827.

[14] T. L. Towner, C. L. Muñoz, A Long Story Short: An Analysis of Instagram Stories during the 2020 Campaigns, Journal of Political Marketing (2022) 1–14. URL: https://doi.org/10.1080/15377857.2022.2099579. doi:10.1080/15377857.2022.2099579.

[15] L. Manovich, Cultural Analytics, MIT Press, 2020.

[16] C. Baden, C. Pipal, M. Schoonvelde, M. A. C. G. van der Velden, Three Gaps in Computational Text Analysis Methods for Social Sciences: A Research Agenda, Communication methods and measures 16 (2022) 1–18. URL: https://doi.org/10.1080/19312458.2021.2015574. doi:10.1080/19312458.2021.2015574.

[17] D. V. Shah, J. N. Cappella, W. R. Neuman, Big Data, Digital Media, and Computational Social Science: Possibilities and Perils, The Annals of the American Academy of Political and Social Science 659 (2015) 6–13. URL: https://doi.org/10.1177/0002716215572084. doi:10.1177/0002716215572084.

[18] T. Araujo, I. Lock, B. van de Velde, Automated Visual Content Analysis (AVCA) in Communication Research: A Protocol for Large Scale Image Classification with Pre-Trained Computer Vision Models, Communication methods and measures 14 (2020) 239–265. URL: https://doi.org/10.1080/19312458.2020.1810648. doi:10.1080/19312458.2020.1810648.

[19] E. Rodina, D. Dligach, Dictator's Instagram: personal and political narratives in a Chechen leader's social network, Caucasus survey 7 (2019) 95–109. URL: https://www.schoeningh.de/downloadpdf/journals/casu/7/2/article-p95_1.pdf. doi:10.1080/23761199.2019.1567145.

[20] V. Murashka, J. Liu, Y. Peng, Fitspiration on Instagram: Identifying Topic Clusters in User Comments to Posts with Objectification Features, Health communication 36 (2021) 1537–1548. URL: http://dx.doi.org/10.1080/10410236.2020.1773702. doi:10.1080/10410236.2020.1773702.

[21] S. Kim, H.-W. Lim, S.-Y. Chung, How South Korean Internet users experienced the impacts of the COVID-19 pandemic: discourse on Instagram, Humanities and Social Sciences Communications 9 (2022) 1–12. URL: https://www.nature.com/articles/s41599-022-01087-7. doi:10.1057/s41599-022-01087-7.

[22] S. Muralidhara, M. J. Paul, #Healthy Selfies: Exploration of Health Topics on Instagram, JMIR public health and surveillance 4 (2018) e10150. URL: http://dx.doi.org/10.2196/10150. doi:10.2196/10150.

[23] A. Al-Rawi, A. Al-Musalli, A. Fakida, News Values on Instagram: A Comparative Study of International News, Journalism and Media 2 (2021) 305–320. URL: https://www.mdpi.com/2673-5172/2/2/18. doi:10.3390/journalmedia2020018.

[24] M. Bossetta, The Digital Architectures of Social Media: Comparing Political Campaigning on Facebook, Twitter, Instagram, and Snapchat in the 2016 U.S. Election, Journalism & mass communication quarterly 95 (2018) 471–496. URL: https://doi.org/10.1177/1077699018763307. doi:10.1177/1077699018763307.

[25] J. W. Rettberg, Snapchat: Phatic Communication and Ephemeral Social Media, in: J. W. Morris, S. Murray (Eds.), Appified: Culture in the Age of Apps, University of Michigan Press, 2018, pp. 188–195.

[26] M. Amancio, "Put it in your Story": Digital Storytelling in Instagram and Snapchat Stories, Master's thesis, Uppsala University, Disciplinary Domain of Humanities and Social

Sciences, Faculty of Social Sciences, Department of Informatics and Media, 2017. URL: https://www.diva-portal.org/smash/record.jsf?pid=diva2%3A1111663&dswid=-5700.

[27] L. Bainotti, A. Caliandro, A. Gandini, From archive cultures to ephemeral content, and back: Studying Instagram Stories with digital methods, New Media & Society (2020) 1461444820960071. URL: https://doi.org/10.1177/1461444820960071. doi:10.1177/1461444820960071.

[28] J. Vázquez-Herrero, S. Direito-Rebollal, X. López-García, Ephemeral Journalism: News Distribution Through Instagram Stories, Social Media + Society 5 (2019) 2056305119888657. URL: https://doi.org/10.1177/2056305119888657. doi:10.1177/2056305119888657.

[29] B. Li, O. K. M. Scott, M. L. Naraine, B. J. Ruihley, Tell Me a Story: Exploring Elite Female Athletes' Self-Presentation via an Analysis of Instagram Stories, Journal of Interactive Advertising 21 (2021) 108–120. URL: https://doi.org/10.1080/15252019.2020.1837038. doi:10.1080/15252019.2020.1837038.

[30] E. A. Nashmi, D. L. Painter, Oh Snap: Chat Style in the 2016 US Presidential Primaries, Journal of Creative Communications 13 (2018) 17–33. URL: https://doi.org/10.1177/0973258617743619. doi:10.1177/0973258617743619.

[31] D. M. Blei, Topic Modeling and Digital Humanities, Journal of Digital Humanities 2 (2012). URL: http://journalofdigitalhumanities.org/2-1/topic-modeling-and-digital-humanities-by-david-m-blei/.

[32] D. M. Blei, Probabilistic Topic Models, Communications of the ACM 55 (2012) 77–84. URL: https://cacm.acm.org/magazines/2012/4/147361-probabilistic-topic-models/fulltext.

[33] M. Grootendorst, BERTopic: Neural topic modeling with a class-based TF-IDF procedure (2022). URL: http://arxiv.org/abs/2203.05794. arXiv:2203.05794.

[34] R. Egger, J. Yu, A Topic Modeling Comparison Between LDA, NMF, Top2Vec, and BERTopic to Demystify Twitter Posts, Frontiers in sociology 7 (2022) 886498. URL: http://dx.doi.org/10.3389/fsoc.2022.886498. doi:10.3389/fsoc.2022.886498.

[35] E. Grave, P. Bojanowski, P. Gupta, A. Joulin, T. Mikolov, Learning Word Vectors for 157 Languages, in: Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018), 2018, pp. 3483–3887.

[36] K. Liebhart, P. Bernhardt, Political storytelling on Instagram: Key aspects of Alexander Van der Bellen's successful 2016 presidential election campaign, Media and communication 5 (2017) 15–25. URL: https://www.cogitatiopress.com/mediaandcommunication/article/view/1062. doi:10.17645/mac.v5i4.1062.

[37] A. Argyrou, S. Giannoulakis, N. Tsapatsoulis, Topic modelling on Instagram hashtags: An alternative way to Automatic Image Annotation?, in: 2018 13th International Workshop on Semantic and Social Media Adaptation and Personalization (SMAP), 2018, pp. 61–67. URL: http://dx.doi.org/10.1109/SMAP.2018.8501887. doi:10.1109/SMAP.2018.8501887.

[38] 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: http://arxiv.org/abs/2103.00020. arXiv:2103.00020.

[39] L. Hemphill, A. M. Schöpke-Gonzalez, Two Computational Models for Analyzing Political Attention in Social Media, Proceedings of the International AAAI Conference on Web and Social Media 14 (2020) 260–271.

Building and Serving the Queerlit Thesaurus as Linked Open Data

Arild Matsson[1], Olov Kriström[1]

[1]The University of Gothenburg, Box 100, 405 30 Gothenburg, Sweden

Abstract

This paper describes the creation of the Queer Literature Indexing Thesaurus (QLIT) as well as the digital infrastructure supporting the workflow for editing and publishing it. The purpose of QLIT is to adequately catalogue Swedish fiction with LGBTQI themes. It is continually edited in plain-text RDF and automatically processed for correctness and storage. Finally, it is published online as Linked Open Data and used with external systems. The technical approach relies on scripts and applications developed ad hoc, rather than existing solutions. Code is available on https://github.com/gu-gridh/queerlit-terms

Keywords

thesaurus, queer, Swedish, Resource Description Framework (RDF), Linked Open Data (LOD)

1. Background

1.1. Queerlit

The Queerlit project aims to identify and index Swedish fiction with LGBTQI themes from the 7th century and onwards. The process of indexing involves assessing a single literary work and choosing suitable subject headings which are then added to the bibliographical record of the work.

Subject headings are organized in a thesaurus, a controlled vocabulary with a hierarchical structure. Subject headings are arranged as broader, more general concepts, and narrower, more specific concepts. In this paper, the term subject heading itself is sometimes also referred to as concept or term.

As a part of the Queerlit project, librarians and literary scholars are together constructing a new thesaurus: *Queer Literature Indexing Thesaurus* (QLIT). This is because existing applicable thesauri are deemed unsuitable for this bibliography, as they often reflect a majority-population (e.g. heterosexual, cisgender) system of concepts. As such, they fail to express the topics and nuances that scholars within queer (Swedish) literary studies would require to appropriately orient themselves within the bibliography. [1, 2, 3, 4, 5]

As an example, in the major Swedish thesaurus Svenska ämnesord (SAO), there is only a single term relating to intersex people (*Intersexualism*). The current version of QLIT, on the

DHNB2023 | Sustainability: Environment - Community - Data. The 7th Digital Humanities in the Nordic and Baltic Countries Conference. Oslo – Stavanger – Bergen, Norway. March 8–10, 2023.

arild.matsson@gu.se (A. Matsson); olov.kristrom@lir.gu.se (O. Kriström)
https://spraakbanken.gu.se/en/about/staff/arild (A. Matsson)
0000-0002-0461-4718 (A. Matsson)

DHNB Publications, DHNB2023 Conference Proceedings, https://journals.uio.no/dhnbpub/issue/view/875

other hand, provides 22 different terms relating to intersex people.

The project runs in the years 2021–2023 and is further described on https://www.gu.se/en/research/queerlit-data-base. As of May, 2023, the project has identified and included around 1700 literary works in its bibliography, and indexed around 950 of these using mainly QLIT.

1.2. Libris XL

The National Library of Sweden (Kungliga biblioteket, KB) keeps bibliographical records (as well as physical copies) of literature published in Sweden. As a participant in the Queerlit project, they ensure the indexing work is being performed and recorded in accordance with national standards in the Libris XL cataloging system. Libris XL uses the BIBFRAME data framework [6, 7] which builds on well established global standards for data interoperability, supported also by the Library of Congress. It is implemented with Linked Open Data (LOD). Through Libris XL, other libraries can get access to the QLIT-enriched bibliographical records and implement them in their local catalogs.

As a step in the ongoing development of the cataloging system, KB recommended for the purpose of the Queerlit thesaurus that its content be defined externally to Libris XL. LOD then provides mechanisms to import this structured data into Libris XL. The requirement upon the Queerlit project is then to expose the subject headings as LOD-compatible data, specifically in the Resource Description Format (RDF).

1.3. Homosaurus

QLIT is based on the Homosaurus, a large thesaurus with a similar purpose [8]. The Homosaurus is created in an international or US-American context, with labels in English, so extensive work was put into adapting it to Swedish conditions.

2. Building the thesaurus

The thesaurus has been built in two phases. In the first, it was created based on the Homosaurus, with translation and broad adjustments. The workflow was then reshaped after the development of the data processing scripts, described in Section 3, and the web server and frontend application, described in Section 4.2. All project members could now easily explore the terms and their hierarchy, give feedback and engage in cataloging work. Thus, in the second phase, smaller adjustments have been made continually.

2.1. Thesaurus creation

Choosing the Homosaurus as the basis of constructing QLIT was based on the fact that it is the largest LGBTQ-themed thesaurus, with both institutional users such as The National Archives in the UK and Duke University Library, and community resources like the Australian Queer Archives and the German QueerSearch. In 2019 it went through a major revision, re-launching as a linked data vocabulary with updated terminology making it more explicitly inclusive of queer and trans topics, which suited the aims of Queerlit.

Since the Queerlit project focuses on fiction, and the Homosaurus is designed as a more general thesaurus, applicable for both fiction and non-fiction, a selection was made of the Homosaurus terms used for indexing fiction at two major institutions using the thesaurus – IHLIA LGBTI Heritage and The Transgender Digital Archives. This selection resulted in a subset of 500 terms (out of the total 1647 terms in Homosarus v2.2). Broader and narrower terms to this selection were included, resulting in a list of 1146 terms which were downloaded as Turtle RDF files [9] and formed the basis for the translation and adaptation.

Initially the work of translating was done manually on a print-out of the list of selected terms, but the Turtle files were soon imported to Visual Studio Code, with an RDF plugin, for text editing. The translations were then made directly in the Turtle files. A benefit of keeping the print-out, however, was that it proved handy as a perspicuous reference for quickly looking up terms.

A benefit of using the Turtle files from the Homosaurus as templates for creating the files for QLIT was that mapping exact and close matches between the two thesauri was easy – a script creating `skos:exactMatch` values from the Homosaurus identifiers was made, and in a case-by-case review some were revised to `skos:closeMatch`. A mapping to the Library of Congress Subject Headings (LCSH) thesaurus was also included with the Turtle files from the Homosaurus – these were also subject to case-by-case review, not only in order to adjust exact matches to close matches, but also in order to correct some misleading matches (mainly having to do with homonyms, where for instance the Homosaurus had an exact match between the term *Bears*, denoting members of a gay subculture, and the LCSH term *Bears*, denoting members of the mammalian ursidae family). These corrections were shared with the Homosaurus editorial board for them to update.

Mappings were also made to the two main Swedish library thesauri, Svenska ämnesord (SAO) and Barnämnesord (Barn). Based on the experience of faulty homonymic matches from the Homosaurus, this mapping was done manually rather than by an automated process.

Further editorial work was done in deciding what terms needed to be removed, adapted or added based on cultural context. For example terms relating to the anglophone distinction of sex/gender were merged in the broad Swedish concept of *Kön*, with narrower terms making distinctions of biological, social, legal and subjective aspects, and a term for Sami LGBTQI people was added as a narrower term in relation to *Indigenous LGBTQI people*.

During the main phase of translating and revising, the identifiers in the Turtle files were handled as plain text, written in camel case (e.g. `dcterms:identifier` *äldreHBTQIPersoner* for the preferred term *Äldre HBTQI-personer* ("Older LGBTQI people")). This facilitated overview and ready comprehension, but in order to safeguard future term changes, these were substituted with randomized, non-semantic identifiers before launch (e.g. `dcterms:identifier` "eg84dq15" for the preferred term *Äldre HBTQI-personer*).

Files were kept on Sharepoint and shared with other project members. A more easily navigated visual representation (see Section 4.2) was launched on a website for use by the project members in February 2022, making collaborative input easier. This also facilitated workshops with groups of end users, who could be presented with the thesaurus in a preliminary design. In October 2022 the thesaurus section of the webpage was made public, allowing an even broader public the possibility of leaving feedback, both regarding the structure and content of the thesaurus and its visual design.

2.2. Continual editing

When the thesaurus was implemented as linked data in the cataloging platform of Libris XL, a second phase of work was initiated. Based on the experiences of indexing the literature revisions were suggested and ideas for further additions were made by the project members. Uploads of revised versions to Libris XL have been made on a semi-regular basis, depending on the rate of changes made.

At the moment of writing QLIT contains 880 terms.

2.3. RDF data model

The RDF data model for QLIT is based on the existing model of the Homosaurus, from which the class `skos:Concept` and the following properties were transposed:

- `dcterm:identifier`
- `dcterm:issued`
- `dcterm:modified`
- `skos:altLabel`
- `skos:broader`
- `skos:closeMatch`
- `skos:exactmatch`
- `skos:inScheme`
- `skos:narrower`
- `skos:prefLabel`
- `skos:related`

After dialogue with KB the property `rdfs:comment`, which is used in the Homosaurus, was not implemented in QLIT, where we instead integrated the property `skos:scopeNote`. The reason for this was that the praxis of KB is to treat `skos:scopeNote` as a general field for both definition and usage guiding, and use of `rdfs:comment` would have been mapped to this in the Libris XL environment.

Further additions were made with the use of the properties `skos:hiddenLabel`, in order to allow common misspellings from end users of the search interface, and `skos:Collection`, as a way to be able to visually organize the thesaurus in a more browsable design, where the (currently) 143 top terms are grouped into 9 thematic categories.

A diagram of the current data model is seen in Figure 1.

3. Data processing

The RDF data created by hand is not guaranteed to be complete or entirely correct. Therefore, the edited Turtle files are synced from the Sharepoint area and input to a script for automatic processing. The script, written in Python, can effortlessly be repeated for each iteration of changes to be published. The steps of the script are described in this section.

The script, named `build.py`, is available in the *queerlit-terms* repository on GitHub: https://github.com/gu-gridh/queerlit-terms.

3.1. Input validation

The first step of the script is to read each input file, validate its content and then add it to a compound RDF graph representing the full thesaurus. This validation step primarily makes sure that the Turtle syntax is valid.

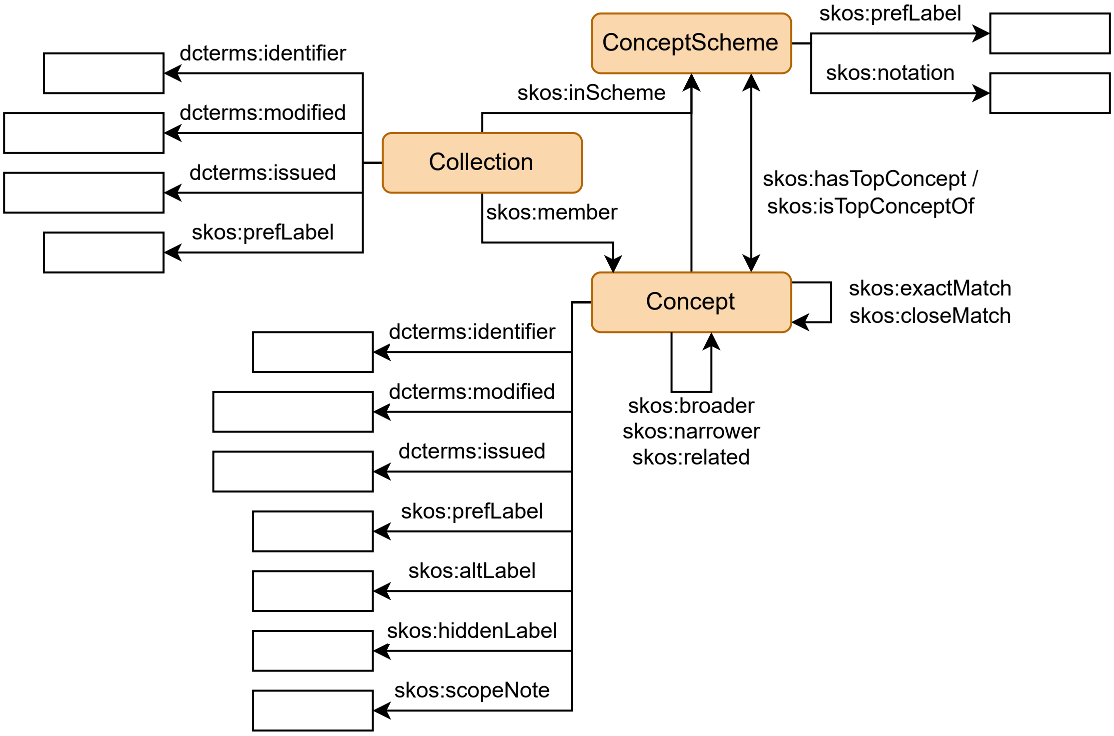


Figure 1: The RDF data model used in QLIT.

After all input files are read and added to the same graph, another validation step checks that there are no relationships to missing nodes, and that the `skos:identifier` value of each term matches its URI.

3.2. Generating randomized identifiers

Each term URI is composed of the thesaurus URI and a term identifier. Early on, identifiers were based on the term label (as mentioned previously, e.g. "äldreHBTQIPersoner" for the term labeled *Äldre HBTQI-personer*). This has the advantage that the term is easily identifiable by human readers. However, labels may be subject to change, while identifiers should remain constant. Thus, we switched to creating identifiers as randomized strings.

To generate identifiers, the Python library StringGenerator[10] was used with a pattern of four lowercase letters and four numbers, for instance: "mg27td65". Resulting identifiers have no meaningful connection to the terms to which they are assigned. The pattern allows 4.5 billion different combinations, equivalent to an entropy of approximately 32 bits. For the unlikely event that a newly generated identifier will already exist for another term, an automated check is in place to retry generation upon such collision.

3.3. Inferring relationships and properties

After loading input files into a unified graph, some trivial changes are made automatically. Most importantly, the `skos:ConceptScheme` node representing the QLIT thesaurus is created, and relationships between it and the terms are added.

Relationships within the thesaurus – `skos:broader`, `skos:narrower` and `skos:related` – are automatically inferred. This means that if *A* has a `skos:broader` relationship to *B*, the script ensures that there is also a `skos:narrower` relation from *B* to *A*.

The SKOS standard leaves some freedom as to how the `skos:topConceptOf` property and its counterpart `skos:hasTopConcept` are applied [11]. We simply define that a *top concept* is any term that does not have any `skos:broader` relationships to another term. The scripts sets these properties accordingly.

3.4. Detecting changes

The result after merging, validating and inferring data is automatically compared to the previous version of the thesaurus. For any added terms, the `dcterms:issued` property is set to the current time. At release-time, the `dcterms:modified` property was also set to the same time. From that point on, for any terms with changes, the `dcterms:modified` property is updated to the current time. This includes both terms in the case where a relationship is added or removed.

```
$ python3 build.py
Parsing 889 files...
Parsed 889 files
Creating new identifiers...
New id cb53to20 for latinesHBTQI
Completing relations...
Checking changes...
Changes for ug69ck69 in narrower
Changes for aa40zb14 in related
Changes for nj44aa03 in narrower
Changes for ho87ol42 in narrower
4 changed, 1 new, 1 removed
Writing 889 terms...
Wrote qlit.nt
```

Listing 1: Sample output from the processing script, summarizing changes made since the previous iteration of the data.

3.5. Version control

At this stage, the RDF graph still exists only transiently in the computer memory. In order to be stored and transferred, it is serialized into the N-Triples format [12], in which triples are output in full and separated by newlines. This output is, in turn, sorted alphabetically.

The sorted N-Triples dump is committed to the *queerlit-terms* git repository. Version control provides long-term storage and facilitates collaboration between developers. Additionally, Git provides a convenient mechanism to diff (i.e. compare) the new output against the previous iteration. We do this routinely as a quick sanity check before committing the output.

There are several serialization formats for RDF, some of which are generally preferred over N-Triples as they are smaller in file size or easier for humans to read. However, they are not quite as easy to sort. Without sorting, the diffs against previous iterations can be less easily interpretable.

3.6. Version number

Versioning of the thesaurus can be considered on two levels. First, the explicit version number is fixed to "v1", and this is part of the thesaurus URI. This is not actually expected to change, but it does provide room for a large-scale re-instantiation of the thesaurus, should the need appear in the future. Second, a small-scale "version" of the thesaurus is implicit in the `dcterms:modified` term property.

Furthermore, the publicly available Git commit history shows in greater detail what changes have been made, but this is more of a by-product, not intended as part of the thesaurus per se.

```
@base <https://queerlit.dh.gu.se/qlit/v1/> .
@prefix dcterms: <http://purl.org/dc/terms/> .
@prefix skos: <http://www.w3.org/2004/02/skos/core#> .
@prefix xsd: <http://www.w3.org/2001/XMLSchema#> .

<rf38ex77> a skos:Concept ;
    dcterms:identifier "rf38ex77" ;
    dcterms:issued "2022-05-19T10:12:00"^^xsd:dateTime ;
    dcterms:modified "2022-12-07T08:49:58"^^xsd:dateTime ;
    skos:broader <iu41ao66>,
        <nf03ub67> ;
    skos:exactMatch
     ↪  <http://id.loc.gov/authorities/subjects/sh85061777> ;
    skos:inScheme <https://queerlit.dh.gu.se/qlit/v1> ;
    skos:narrower <pn00dm58>,
        <yl06dh66> ;
    skos:prefLabel "Homosexuella par" ;
    skos:related <vu78ao12> ;
    skos:scopeNote "Används för skildringar av parrelationer mellan
     ↪  homosexuella." .
```

Listing 2: The RDF (Turtle) representation of the term *Homosexuella par* ("Homosexual couples") as a result after automatic processing.

Table 1
RDF routes

Route name	URL	Content (RDF/Turtle)
RDF All	https://queerlit.dh.gu.se/qlit/v1/	The full QLIT thesaurus
RDF One	https://queerlit.dh.gu.se/qlit/v1/<identifier>	A single QLIT subject heading

4. Web server

A lightweight web server application was created to serve the data online for two purposes. Firstly, the full RDF data needs to be made available for importing into Libris XL. Secondly, the Queerlit website (https://queerlit.dh.gu.se/) needs to query and navigate the data. These parts are detailed in each of the following two subsections.

The application loads the version-controlled data dump and serves a few different routes under the URL "https://queerlit.dh.gu.se/qlit/v1/". It is written in Python, like the data processing script, and it lives in the same code repository. It uses the web server library Flask.

4.1. RDF server

The URIs for the `skos:ConceptScheme` node representing the QLIT thesaurus, as well as for all the `skos:Concept` nodes representing QLIT subject headings, are designed to be valid URLs, resolving to their RDF representations. The server application is then configured to serve under the first part of the URL ("https://queerlit.dh.gu.se/qlit/v1/") and designed to resolve the last part of the URL (the empty string or a subject heading identifier) to the `skos:ConceptScheme` or a `skos:Concept`.

While N-Triples was used as output format for the version-controlled data dump, Turtle is chosen in this case for its readability. The *RDF All* route is in principle just a re-serialization of the data dump into Turtle.

The RDF content of these routes is intended as the canonical, "source of truth" variant of the QLIT thesaurus to the outside world.

4.2. JSON server for frontend application

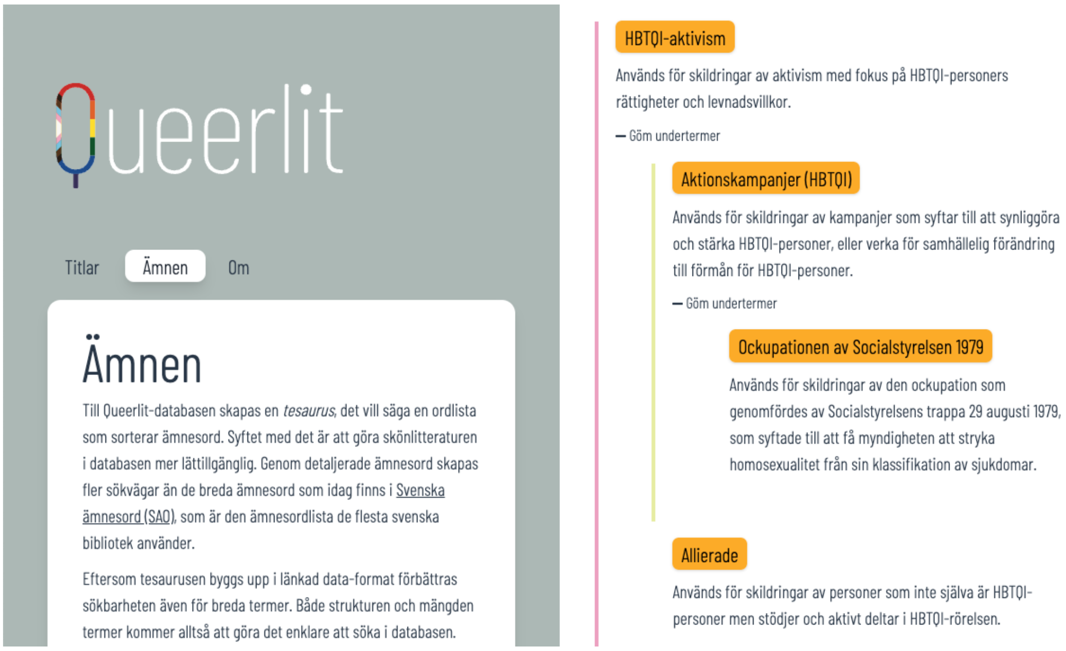
Another goal of the Queerlit project is creating a website for searching the Queerlit bibliography and browsing the QLIT thesaurus. It is a frontend-only application which runs in the user's web browser and fetches data primarily from the Libris XL web API.

Thesaurus data could also be fetched from Libris XL, but is instead fetched from the QLIT server application for two reasons. First, the QLIT data in Libris XL is in theory only a cached version of the canonical data served by us. Second, we can enrich the data with labels from matching Homosaurus terms. Third, the output format is designed to be easy to use in the frontend application code.

Each route corresponds to a method of traversing the RDF graph and identifying a set of terms. Each term is translated to a Python object with properties corresponding roughly but not exactly to RDF triples connected to the term.

Table 2
JSON routes

Route name	Content (JSON)
All	List of all terms in no specific order
One	One term
Autocomplete	List of terms matching a given search string
Roots	List of all terms having no `skos:broader` relation
Children	List of all terms having a `skos:broader` relation to a given term
Parents	List of all terms having a `skos:narrower` relation to a given term
Related	List of all terms having `skos:related` relation to a given term



Ämnen

Till Queerlit-databasen skapas en *tesaurus*, det vill säga en ordlista som sorterar ämnesord. Syftet med det är att göra skönlitteraturen i databasen mer lättillgänglig. Genom detaljerade ämnesord skapas fler sökvägar än de breda ämnesord som idag finns i Svenska ämnesord (SAO), som är den ämnesordlista de flesta svenska bibliotek använder.

Eftersom tesaurusen byggs upp i länkad data-format förbättras sökbarheten även för breda termer. Både strukturen och mängden termer kommer alltså att göra det enklare att söka i databasen.

HBTQI-aktivism

Används för skildringar av aktivism med fokus på HBTQI-personers rättigheter och levnadsvillkor.

— Göm undertermer

Aktionskampanjer (HBTQI)

Används för skildringar av kampanjer som syftar till att synliggöra och stärka HBTQI-personer, eller verka för samhällelig förändring till förmån för HBTQI-personer.

— Göm undertermer

Ockupationen av Socialstyrelsen 1979

Används för skildringar av den ockupation som genomfördes av Socialstyrelsens trappa 29 augusti 1979, som syftade till att få myndigheten att stryka homosexualitet från sin klassifikation av sjukdomar.

Allierade

Används för skildringar av personer som inte själva är HBTQI-personer men stödjer och aktivt deltar i HBTQI-rörelsen.

Figure 2: A screenshot from the thesaurus section of the Queerlit website (https://queerlit.dh.gu.se/subjects).

The Homosaurus is loaded as a separate RDF graph. For QLIT terms with an `skos:exactMatch` relation to a Homosaurus term, the term is looked up and the `skos:prefLabel` and `skos:altLabel` values are added to the Python object.

The list of objects (or single object, in the case of the *One* route) are serialized to the JSON format, chosen because of its easy and widespread use in frontend applications.

The API is openly accessible in practice, but not intended for this purpose. It is not publicly documented, and changes are not announced.

5. Discussion

5.1. Custom implementation

As detailed in this paper, the process has involved custom implementation for data validation, processing and web publication. Writing our own code, rather than employing existing available software, has enabled solving specific problems without adding irrelevant cruft. However, as the project evolves, there is an increased probability that problems get more complex and alike those that existing software is solving.

Some solutions which are already openly available, and which could perhaps have been used instead of custom code, are:

- Editor applications specialized for RDF (e.g. Protégé [13]). We started working with a non-specialized text editor because it was easier to get started. Investing time in learning a new application might have saved us time in the long run, but we did not consider it until later in the process. Any mistakes that would have been avoided with a specialized editor (e.g. relations to non-existing nodes) were detected and corrected in the input validation process.
- Applications for publishing RDF online (e.g. Skosmos [14]). We considered existing applications, but writing a custom application was not a very time-consuming alternative. It also gave us the possibility to write endpoints suitable for use with the search website.
- Applications for visualization of the thesaurus. We wanted to integrate this visually with the search website, which is also designed and implemented as part of the research project. We made the judgment that employing existing visualization tools with their own design would break the user experience, and overriding their design would be time-consuming.

References

[1] J. Hansson, Klassifikation, bibliotek och samhälle : en kritisk hermeneutisk studie av "Klassifikationssystem för svenska bibliotek", Skrifter från Valfrid, 19, 1999.

[2] G. Campbell, Queer Theory and the Creation of Contextual Subject Access Tools for Gay and Lesbian Communities, KO KNOWLEDGE ORGANIZATION 27 (2000) 122–131. URL: https://www.nomos-elibrary.de/10.5771/0943-7444-2000-3-122/. doi:10.5771/0943-7444-2000-3-122, publisher: Nomos Verlagsgesellschaft mbH & Co. KG.

[3] H. A. Olson, Power to name: locating the subject representation in libraries, Dordrecht ; Boston, Mass. Kluwer, 2002.

[4] J. Samuelsson, På väg från ingenstans: kritik och emancipation av kunskapsorganisation för feministisk forskning, Akademiska avhandlingar vid Sociologiska institutionen / Umeå universitet, 53, Umeå universitet, Diss. Umeå, 2008. URL: urn:nbn:se:umu:diva-1822.

[5] J. Bates, J. Rowley, Social reproduction and exclusion in subject indexing: A comparison of public library OPACs and LibraryThing folksonomy, Journal of documentation 67 (2011) 431–448. Place: BINGLEY Publisher: EMERALD GROUP PUBLISHING.

[6] S. McCallum, BIBFRAME Development, JLIS.it 8 (2017) 71–85. URL: https://jlis.fupress.net/index.php/jlis/article/view/127. doi:10.4403/jlis.it-12415, number: 3.

[7] Vad är Libris och XL?, ????. URL: https://libris.kb.se/katalogisering/about.

[8] About, ????. URL: https://homosaurus.org/about.

[9] D. Beckett, T. Berners-Lee, E. Prud'hommeaux, G. Carothers, RDF 1.1 Turtle, 2014. URL: https://www.w3.org/TR/turtle/.

[10] P. Wolf, StringGenerator: Generate randomized strings of characters using a template, 2021. URL: https://github.com/paul-wolf/strgen.

[11] A. Miles, S. Bechhofer, SKOS Simple Knowledge Organization System Reference, 2009. URL: https://www.w3.org/TR/skos-reference/#schemes.

[12] D. Beckett, RDF 1.1 N-Triples, 2014. URL: https://www.w3.org/TR/n-triples/.

[13] J. H. Gennari, M. A. Musen, R. W. Fergerson, W. E. Grosso, M. Crubézy, H. Eriksson, N. F. Noy, S. W. Tu, The evolution of Protégé: an environment for knowledge-based systems development, International journal of human-computer studies 58 (2003) 89–123. doi:10.1016/S1071-5819(02)00127-1, place: LONDON Publisher: Elsevier Ltd.

[14] O. Suominen, H. Ylikotila, S. Pessala, M. Lappalainen, M. Frosterus, J. Tuominen, T. Baker, C. Caracciolo, A. Retterath, Publishing SKOS vocabularies with Skosmos (2015).

Community and Interoperability at the Core of Sustaining Image Archives

Ulrike Felsing[1,3], Peter Fornaro[2], Max Frischknecht[1,3] and Julien Antoine Raemy[2,4]

[1]*HKB - Bern Academy of the Arts, Bern University of Applied Sciences, Fellerstrasse 11, 3027 Bern, Switzerland*
[2]*Digital Humanities Lab, University of Basel, Spalenberg 65, 4051 Basel, Switzerland*
[3]*Walter Benjamin Kolleg, University of Bern, Muesmattstrasse 45, 3012 Bern, Switzerland*
[4]*DaSCH - Swiss National Data and Service Center for the Humanities, Gewerbestrasse 24, 4123 Allschwil, Switzerland*

Abstract

In our paper, we discuss how the digital domain extends the sustainability of analogue archives through communication with the public. Our interdisciplinary research project "Participatory Knowledge Practices in Analogue and Digital Image Archives" (PIA) is funded by the Swiss National Science Foundation (2021–2025) and developed in cooperation with the photographic archives of the Swiss Society for Folklore Studies (SSFS). It aims to increase the use of image-based research data by developing participatory tools and deploying shared application programming interfaces (APIs) such as standards that adhere to the Linked Open Usable Data (LOUD) design principles. By involving the public, the project aims to increase the overall use of image-based research data. This makes data more sustainable in interaction with the analogue archive and increases the attractiveness of digital infrastructures.

Keywords

Citizen Science, Cultural Heritage, Digital Infrastructure, Interoperability, Linked Open Usable Data, Participatory Design, Sustainability

1. Introduction

We are engaged in an ongoing research project that explores open collaboration in the context of photographic archives. Our objective is to develop interactive tools and interfaces that foster the utilisation of image-based research data through active participation. By facilitating engagement from various communities, we aim to encourage the collective production of knowledge and democratise decision-making processes within archival settings. In our paper, we investigate how an interconnected and usable digital realm extends the sustainability of analogue archives through interaction and public involvement. We also emphasise the importance of standards in enabling data sharing and improving the resilience of cultural heritage data.

DHNB2023 | Sustainability: Environment - Community - Data. The 7th Digital Humanities in the Nordic and Baltic Countries Conference. Oslo – Stavanger – Bergen, Norway. March 8–10, 2023.

✉ ulrike.felsing@hkb.bfh.ch (U. Felsing); peter.fornaro@unibas.ch (P. Fornaro); max.frischknecht@hkb.bhf.ch (M. Frischknecht); julien.raemy@unibas.ch (J. A. Raemy)
🌐 http://ulrikefelsing.ch/ (U. Felsing); https://maxfrischknecht.ch/ (M. Frischknecht); https://julsraemy.ch/ (J. A. Raemy)
iD 0000-0003-3791-3316 (U. Felsing); 0000-0003-1485-4923 (P. Fornaro); 0000-0001-8043-1895 (M. Frischknecht); 0000-0002-4711-5759 (J. A. Raemy)

DHNB Publications, DHNB2023 Conference Proceedings, https://journals.uio.no/dhnbpub/issue/view/875

We first provide a short state of the art in terms of building communities of citizen scientists in section 2. It is followed by section 3 that emphasises on the research project Participatory Knowledge Practices in Analogue and Digital Image Archives (PIA). We then discuss Linked Open Usable Data (LOUD) in section 4 and their communities that develop and maintain shared application programming interfaces (APIs) for semantic interoperability purposes. In sections 5 and 6 we address the participatory design of the graphical user interface (GUI) and the overall technical architecture of PIA. Section 7 outlines suggestions for further research and finally in section 8, we conclude our analysis.

2. Sustaining Archives with Citizen Scientists

An increasing amount of research is being done in open collaboration with a crowd, with some of these projects being understood as Citizen Science which is characterised by openness in terms of participation and thus offers diverse perspectives for engagement within different fields of knowledge. Similar projects include Ajapaik [1] for crowdsourcing additional visual heritage metadata, Corley Explorer [2] for collecting stories, sMapshot [3] for georeferencing images, or Historypin [4] and notreHistoire.ch [5] for sharing local history.

According to Ridge, moderating contributions is one of the bigger challenges. For example, one participant reported feeling frustrated about moderation delays on her posts, which indicates that a lack of timely feedback on a completed task has a detrimental effect. In contrast,

> *"others [. . .] felt that the changing source material they were working with helped them stay motivated."* [6, 131]

These and other issues of curating collaboration will always be the responsibility of the group that initiates the "Call for images". On our platform, however, we must provide the adequate functionalities to do so. This includes, for example, the login and the possibility to report problematic content.

3. Participatory Knowledge Practices in Analogue and Digital Image Archives

Participatory Knowledge Practices in Analogue and Digital Image Archives (PIA)[1] is a four-year research project (2021-2025). The project is led by the University of Basel (Institute for Cultural Anthropology and European Ethnology as well as the Digital Humanities Lab) and the Bern Academy of the Arts (HKB).

Our research is based on three cultural heritage collections from the photographic archives of the Swiss Society for Folklore Studies (SSFS)[2]: one focusing on scientific cartography (Atlas of Swiss Folklore, published from 1950 until 1995), a second from the estate of the photojournalist Ernst Brunner (1901–1979), and a third collection consisting of vernacular photography which was owned by the Kreis Family (1860–1970).

[1] https://about.participatory-archives.ch

[2] https://archiv.sgv-sstp.ch/

- The *Atlas of Swiss Folklore* was commissioned by the SSFS as part of a long-term scientific project. By mapping cultural patterns, the atlas constructs a comprehensive picture of Switzerland based on the links between culture and geography. The data was compiled by both academic professionals and non-academics, called "laymen" in the language of the time. The goal of this extensive survey was to document "Swiss folk culture" in the 1930s and 1940s through questions on a wide range of topics such as everyday behaviour, local laws, festivals, work and trade [7].
- *Ernst Brunner* consists of 48,000 negatives and 20,000 prints. Brunner, a professional photographer, published numerous artistic and documentary photographs on a wide variety of folkloric subjects, primarily in popular magazines e.g. "Swiss Home" and "Swiss Family" [8].
- *Kreis Family* is a private collection of a Basel-based family of physicians and printers that connects many strands of the cultural imagination. It is a typical example of urban bourgeois culture. The collection includes 20,000 loose photographic objects. A quarter of them were organised in 93 photo albums.

In a decisive interdisciplinary approach between digital humanities, cultural anthropology and design research, we work together to connect practical and theoretical issues, especially the user perspectives identified in the workshops and the theoretical findings. We are committed to collaborating with the scientific community and the wider public, facilitating the preservation and dissemination of knowledge, and encouraging users to engage together with their own histories and contemporary practices

Overall, our main goal is to increase the use of the images and their metadata of the three collections. For this purpose, we develop digital tools that support contextualising, linking and contrasting images. By fostering exchange and collaboration in digital communities, we contribute to strengthening the photographic archives of the SSFS, and in order to achieve this we rely as much on the development of APIs maintained by cultural heritage practitioners as on the creation of a GUI following participatory design guidelines.

4. Semantic Interoperability through Linked Open Usable Data

Collaboratively designed standards build the base by which individuals and institutions can participate by having unmitigated access to the data. We consider, in particular, that Linked Open Usable Data (LOUD) specifications [9] are adequate not only in the cultural heritage field but more broadly for all participatory efforts within Citizen Science practices as it is an approach to serialise and expose data for different target groups.

> *"One of the first intentions of LOUD is to provide access to the data for both academic and software developers. An appropriate balance must be found between the requirements of data completeness and accuracy, driven by the ontological construct, and the pragmatic concerns of scalability and usability".* [10]

Similar to Tim-Berners Lee's 5-star deployment scheme for Open Data[3], five design principles

[3] https://5stardata.info/

underpin LOUD[4]:

- **The right Abstraction for the audience**: use cases rather than ontological purity should be favoured to determine the level of interoperability.
- **Few Barriers to entry**: the data, and the underlying model, must be easy to leverage. Having such systems in place will encourage more people to actively use them.
- **Comprehensible by introspection**: the data must be largely understandable simply by looking at it, without requiring external help.
- **Documentation with working examples**: comprehensive documentation should be produced to clarify the implementation of the use cases.
- **Few Exceptions, instead many consistent patterns**: patterns should be able to accommodate as few exceptions as possible to avoid adding rules that require the creation of tailor-made leeway fields on a case-by-case basis.

These design principles were inspired by those of the International Image Interoperability Framework (IIIF)[5], a community-driven initiative made up largely of academic and memory institutions. Since 2012, IIIF has developed and maintained shared APIs for representing and annotating digital resources [11] and it has fundamentally shaped how libraries, archives and museums disseminate digital surrogates and digital-born objects over the past few years.

If we consider the relationship between the FAIR data principles (*Findable, Accessible, Interoperable, Reusable*) [12] and LOUD, it appears that the former refers to the environment in which the data is situated, while the latter pertains to the content itself. Breaking down the LOUD

[4] https://linked.art/loud/

[5] https://iiif.io/api/annex/notes/design_principles/

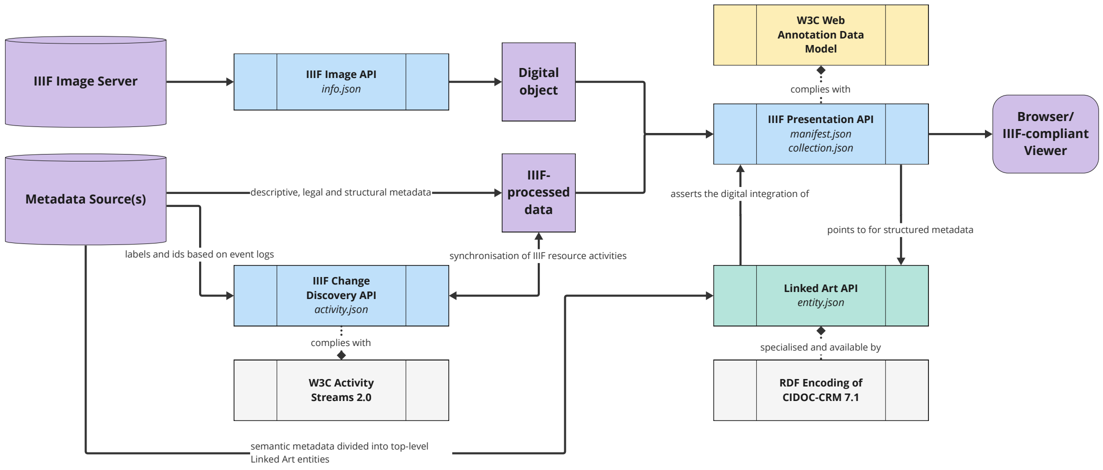

W3C Web Annotation Data Model
IIIF Image Server
IIIF Image API
info.json
Digital object
complies with
IIIF Presentation API
manifest.json
collection.json
Browser/ IIIF-compliant Viewer
Metadata Source(s)
descriptive, legal and structural metadata
IIIF-processed data
asserts the digital integration of
points to for structured metadata
labels and ids based on event logs
synchronisation of IIIF resource activities
IIIF Change Discovery API
activity.json
Linked Art API
entity.json
complies with
specialised and available by
W3C Activity Streams 2.0
RDF Encoding of CIDOC-CRM 7.1
semantic metadata divided into top-level Linked Art entities

Figure 1: Example of a LOUD ecosystem which illustrates how the IIIF APIs (here the Image, Presentation and Change Discovery APIs), the Web Annotation Data Model and the Linked Art API can be integrated in the same environment. This diagram was inspired by the one on https://iiif.io/get-started/how-iiif-works/.

acronym, the terms *Linked*, *Usable* and *(machine-readable) Data* can be seen as characteristics of the data (and its usability once transferred to an environment) [10]. *Open* can correspond somewhat to the principle of reusability outlined in FAIR [13].

Among the standards that adhere to the LOUD design principles are of course those conceived by the IIIF community, especially the IIIF Presentation API 3.0 which has been made with the latest updates of their design principles and for easier integration with JSON-LD 1.1 and other Web standards [14] such the Web Annotation Data Model [15], as well as Linked Art, a Resource Description Framework (RDF) application profile of CIDOC-CRM — a high-level ontology to enable information integration for cultural heritage data [16] — for semantically conveying assertions in a event-based paradigm [17]. All of these specifications were made and created collaboratively. While the Web Annotation Data Model is the result of a World Wide Web Consortium (W3C) working group that had a limited lifespan, the other two are established and open communities with participants mainly from the cultural heritage field and the Digital Humanities.

As shown in Figure 1, these specifications can be implemented separately and in conjunction with each other, since they rely on the same technological foundation as they are serialised in JavaScript Object Notation for Linked Data (JSON-LD), which allows some mapping of ontological constructs into JSON, the lingua franca for most of current web applications. Above all, these standards are predominantly built and maintained from the ground up while adhering to the architecture of the World Wide Web [18], reusing existing standards where possible. For example, the IIIF Change Discovery API relies on the W3C Activity Streams 2.0 standard to describe changes to resources and facilitates crawling to build search indexes [19].

As an exemplary model for other organisations, LUX[6], the Yale Collection Discovery platform, has been upgraded to rely on LOUD standards, notably by leveraging Linked Art for facilitating data sharing across domains, and IIIF for seamless image delivery. It brings together over 40 million cultural heritage records (objects, works, people and organisations, places, concepts, and events) from libraries and museums within Yale University, making it a valuable reference for other institutions. While providing access to Yale's collections, LUX also reconciles its data with external sources like Wikidata. This integration not only allows users to explore a vast range of cultural artefacts and information within Yale's collections but also sets a precedent for other organisations aiming to enhance their own cultural heritage platforms. In short, LOUD could be considered as a pragmatic perspective for presenting or describing as well as sharing cultural heritage data on the Web that responds to the usability requirements of both scholars and developers working jointly through use cases. Thus, we strongly believe that implementing LOUD standards in the PIA digital infrastructure is a way to communicate specific values about sharing and open data practices, as well as to improve the resilience of cultural heritage data.

However, it is worth noting that one of the still largely unexplored areas of study concerning LOUD standards, in particular for the IIIF and Linked Art specifications which are exposed to a greater extent than the Web Annotation Data Model normally baked into other standards, is to carry out an assessment of these APIs in terms of usability factors, precisely to appraise one of the drivers of the perspective[7].

[6] https://lux.collections.yale.edu/

[7] What has already been done within the IIIF community are, for example, usability tests of IIIF-compatible

5. Participatory Design of Graphical User Interface

In this section, we discuss our approach to the participatory design of the graphical user interface. We believe that participatory development is essential in creating helpful and comprehensible GUI's. We understand participatory research as the attempt to involve individuals and groups affected by the respective topic and question as active decision-makers [22]. Such participation can occur at different stages of the interface development process, for example, during conceptualisation, design, collaborative usage or evaluation of the prototypes. For our user research we apply a combination of *Activity Centred* [23] and *Goal Directed* [24] design methods that aim at understanding users goals and motivations in relation to activities and tasks performed in the interface. In doing so, we aim to develop a coherent user experience (UX) and interface principles that facilitate an active engagement beyond the project duration.

In the following, we highlight experiences from an early stage of the interface development where we worked with different humanities scholars from the fields of History and Cultural Anthropology that work with the photographic archives of the SSFS in their projects. The goal of the workshops was to define some of the visual, communicative, and functional requirements specific to humanities scholars for searching and interpreting digital cultural photographic collections. While Humanities scholars are one of the platform's main target groups, we plan to diversify our workshop participants throughout the project duration to involve a broader public.

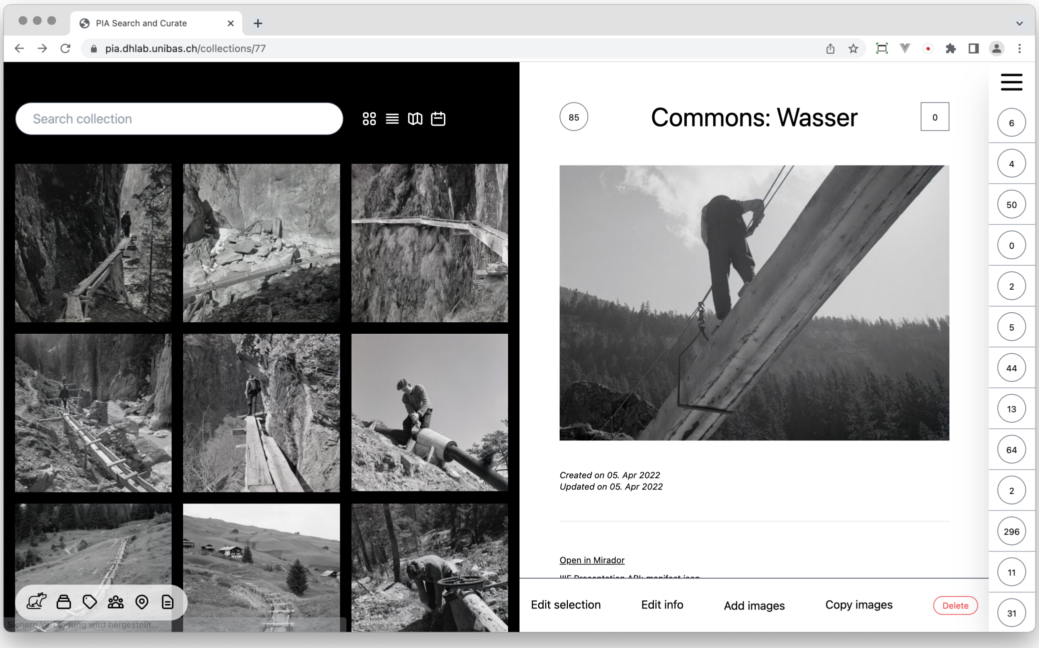
One of the two use cases of the 2022 workshops was the *Images of Swiss Commons* project. It aims to document and explore the mutability and innovation potential of the Swiss commons and the collective action of alpine farmers. This is performed by means of historical and contemporary photographs. Different actors of today's collective will be involved, i.e. current or former officials of civil communities, businesses, alpine cooperatives, etc., as well as local historians.

With the support of the PIA platform, the researchers would like to use Ernst Brunner's photo collection to enter into dialogue with representatives of interested organisations through "Calls for images". The specific communities addressed will be invited to contextualise these historical photographs with their own visual documents on their collective forms of ownership and management.

Photographs are at the heart of the project because they offer non-historians a direct view of the past and a low-threshold access to the diversity of material and immaterial culture. Photographs provide important clues to cultural practices because they can convey implicit knowledge. This makes them valuable starting points for discussions that can be held on the digital platform, but also in real workshops. Hereafter are some of the challenges and questions that were discussed during the workshop:

- How should the comments and discussions with the addressed groups be curated and moderated?
- Is a login necessary or should we provide a low-threshold access to increase the chances of participation?

software [20, 21] or the creation of personas by the IIIF Design Community Group (see https://iiif.io/guides/guides/personas/), but not yet an assessment at the specification level.



Created on 05. Apr 2022
Updated on 05. Apr 2022

Figure 2: One of the first prototypes of the GUI — "Search & Curate" — showing how a collection can be displayed (here from the project "Images of Swiss Commons").

- Should the addressed groups be able to react to specific photographs or do we formulate open "Calls for Images"?

One concrete outcome of the 2022 workshop was a more detailed envisioning of how users can create a sub-collection through our platform (see Figure 2). This also includes the differentiation of a collection into specific themes, for example on the resources water, forest and alpine pasture in *Images of Swiss Commons*.

For the second use case, we collaborated with a member from the *Mensch & Haus*[8] research project that documents the cultural evolution of historic farmhouses and their inhabitants across Switzerland. In this scenario the PIA platform is used to organise and analyse historic photographs from the landscape of Adelboden in Switzerland to support on-sight ethnographic surveys with inhabitants. In the workshop, a user journey (see Figure 3) was developed collaboratively to facilitate discussion and define the researcher's motivation, tasks and challenges in using our GUI.

This collaboration helped us to understand two fundamental aspects that further informed the development of our GUI:

1. Our platform is used in different phases of research and is one of several tools with which researchers work. This highlights the emphasis on user-friendly import and export

[8] https://data.snf.ch/grants/grant/189398

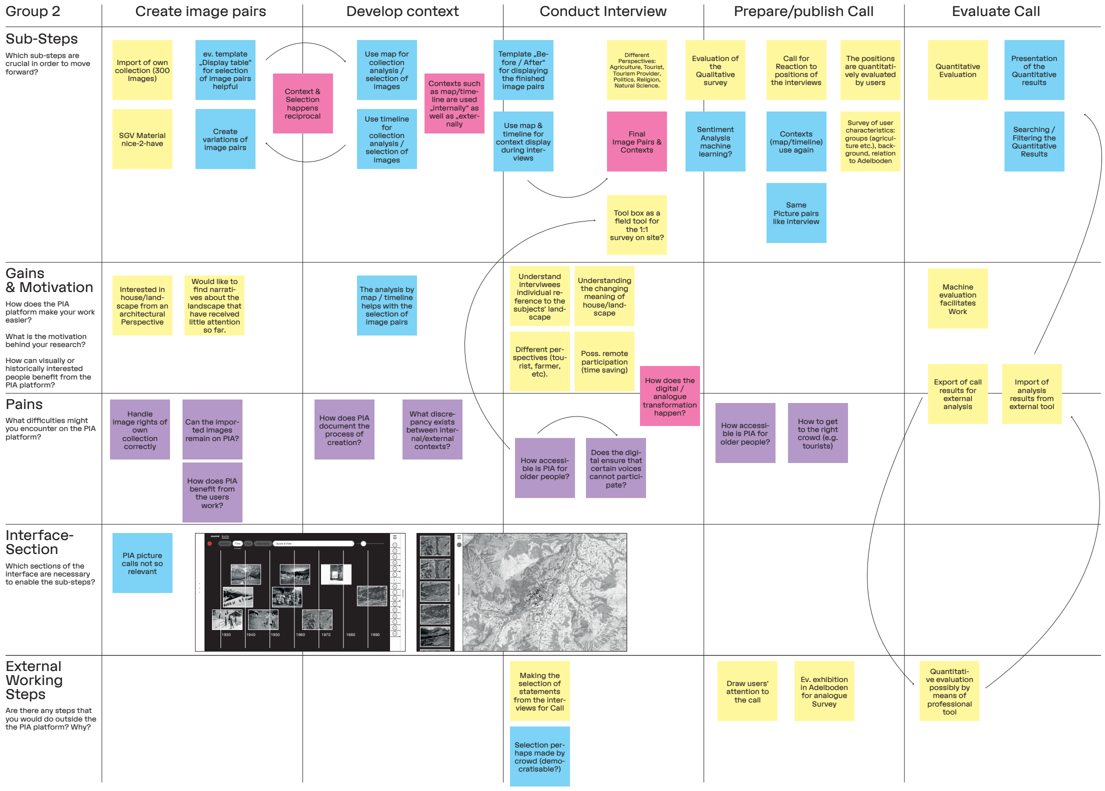

Group 2
Create image pairs
Develop context
Conduct Interview
Prepare/publish Call
Evaluate Call
Sub-Steps
Which sub-steps are crucial in order to move forward?
Import of own collection (300 Images)
SGV Material nice-2-have
ev. template „Display table“ for selection of image pairs helpful
Create variations of image pairs
Context & Selection happens reciprocal
Use map for collection analysis / selection of images
Use timeline for collection analysis / selection of images
Contexts such as map/timeline are used „internally“ as well as „externally
Template „Before / After“ for displaying the finished image pairs
Use map & timeline for context display during interviews
Different Perspectives: Agriculture, Tourist, Tourism Provider, Politics, Religion, Natural Science.
Final Image Pairs & Contexts
Tool box as a field tool for the 1:1 survey on site?
Evaluation of the Qualitative survey
Sentiment Analysis machine learning?
Call for Reaction to positions of the interviews
Contexts (map/timeline) use again
Same Picture pairs like interview
The positions are quantitatively evaluated by users
Survey of user characteristics: groups (agriculture etc.), background, relation to Adelboden
Quantitative Evaluation
Presentation of the Quantitative results
Searching / Filtering the Quantitative Results
Gains & Motivation
How does the PIA platform make your work easier?
What is the motivation behind your research?
How can visually or historically interested people benefit from the PIA platform?
Interested in house/landscape from an architectural Perspective
Would like to find narratives about the landscape that have received little attention so far.
The analysis by map / timeline helps with the selection of image pairs
Understand interviwees individual reference to the subjects‘ landscape
Understanding the changing meaning of house/landscape
Different perspectives (tourist, farmer, etc).
Poss. remote participation (time saving)
How does the digital / analogue transformation happen?
Machine evaluation facilitates Work
Export of call results for external analysis
Import of analysis results from external tool
Pains
What difficulties might you encounter on the PIA platform?
Handle image rights of own collection correctly
Can the imported images remain on PIA?
How does PIA benefit from the users work?
How does PIA document the process of creation?
What discrepancy exists between internal/external contexts?
How accessible is PIA for older people?
Does the digital ensure that certain voices cannot participate?
How accessible is PIA for older people?
How to get to the right crowd (e.g. tourists)
Interface-Section
Which sections of the interface are necessary to enable the sub-steps?
PIA picture calls not so relevant
Journal
Suche
Time
Map
Suche & Filter
1930
1940
1950
1960
1970
1980
1990
External Working Steps
Are there any steps that you would do outside the the PIA platform? Why?
Making the selection of statements from the interviews for Call
Selection perhaps made by crowd (democratisable?)
Draw users‘ attention to the call
Ev. exhibition in Adelboden for analogue Survey
Quantitative evaluation possibly by means of professional tool

Figure 3: The filled out user-journey after the finished workshop. The X-axis describes the five main tasks the researcher will conduct in chronological order. The Y-axis describes motivation, potential problems, and used interface components for each task.

functionalities that allow data to be transferred in and out of the platform in standardised formats such as JSON or CSV. This is in line with the principle of open systems and usable data, which gives users sovereignty over their own data [25]. It further shows that the LOUD design principles described earlier, especially the first two (*The right abstraction for the audience* and *Few barriers to entry*), are relevant for users.

2. As the researcher in this scenario works with a large image corpus of 300 images, it's important that the interface provides the appropriate tools to explore the corpora intuitively. This includes, for example, the ability to group and organise images and to visualise them in different contexts, such as a timeline or a map. This led to the idea of generating multiple views in the interface (see Figure 4). Each group of images can be displayed in different visual contexts, such as a grid, table, map, timeline or network visualisation. Each view provides the user with specific interaction possibilities. For example: While the table provides advanced sorting functionality, the map allows searching for similar images within a certain radius of a given location.

The overall goal of the PIA platform is to allow scholars and citizen scientists to engage with,

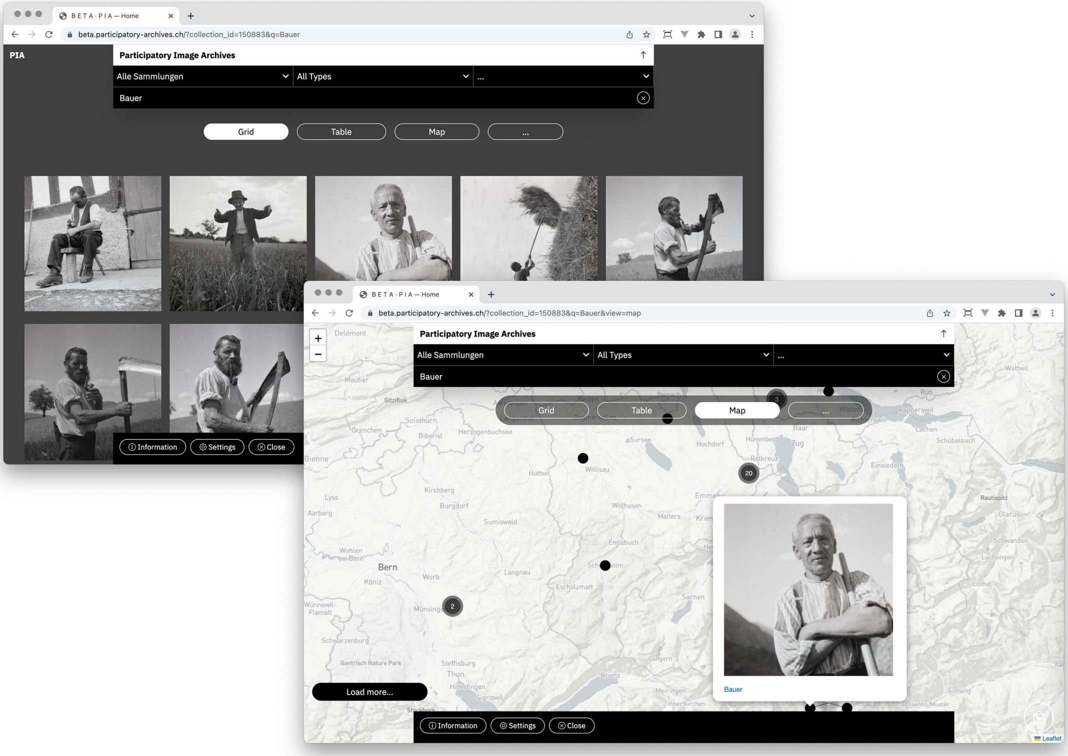

BETA-PIA — Home
beta.participatory-archives.ch/?collection_id=150883&q=Bauer
PIA
Participatory Image Archives
Alle Sammlungen
All Types
...
Bauer
Grid
Table
Map
...
Information
Settings
Close
beta.participatory-archives.ch/?collection_id=150883&q=Bauer&view=map
Bern
Load more...
Bauer
Leaflet

Figure 4: Current prototype of the GUI showing how the same set of images can be displayed in different visual contexts such as a grid or map to support the contextualisation of the data.

edit, enrich and curate data of cultural heritage collections like the SSFS archive. On the one hand, this aims at enabling crowdsourcing in the usual sense by moderated metadata enrichment [6, 26], for example, when residents of an area make comments on historical photographs. On the other hand, the interface offers more far-reaching possibilities of participation by allowing users to create individual sub-collections and launch open "Calls for Images". These user-generated collections are compilations of archival images and documents, as well as personal uploads and descriptions, that can be edited by individual or groups of users.

The challenge here is to find a good balance between control over content and low-threshold access. Related to this are in particular questions around the login-policy of the platform. The first prototypes prioritised easy access and most functionalities didn't require a login. As the platform scales we are gradually introducing login requirements while trying to preserve easy access. For example: logins are not necessary to create collections in the GUI or annotating images in Mirador[9], a IIIF-compliant viewer, as they are temporarily saved in the browser cache. To save collections and annotations in the long term and to edit them in detail, a login is required. Similarly, with "Calls for images", while a login is not required to submit entries, it is necessary for editing and curating user submissions. Other functions are planned with login:

[9] https://projectmirador.org/

editing one's own sub-collection, initiating individual "Calls" as well as editing annotations or receiving appropriate credit for this.

The design and development of the PIA interface is enabled by a flexible software architecture using a headless approach, separating the back-end from the front-end and connecting them through the API. Inspired by the atomic design approach [27], each interface view (e.g. Grid, Table or Map) can be developed as an independent front-end component based on the same data structure, coding conventions and design system. On the one hand, this allows us to implement future views derived from new user requirements, on the other hand, it allows other projects to adopt parts of our codebase into their own. This approach further contributes to the GUI's sustainability beyond the projects duration. It allows to maintain front-end and back-end separately, making it easier, for example, to find suitable partners for further development and make changes to parts of the used technology. The component based architecture enables the exchange of parts of the GUI without having to revise the overall architecture. We are currently discussing different models to ensure the sustainability of the GUI after the end of the project in 2025. We are working together with the SSFS photo archive, which could take over the maintenance. However, the costs incurred would have to be covered by the projects that want to carry out the "Calls for images".

6. PIA Overall Architecture

Within this section, we will first cover the high-level overview of the infrastructure and its environment, and then go into more detail about the generation of LOUD resources.

As mentioned previously, the software architecture is strictly separated into a headless back-end that provides the potential to represent data in complex digital data models, various interfaces for communication, and a front-end framework that is capable of embedding tools for different types of applications.

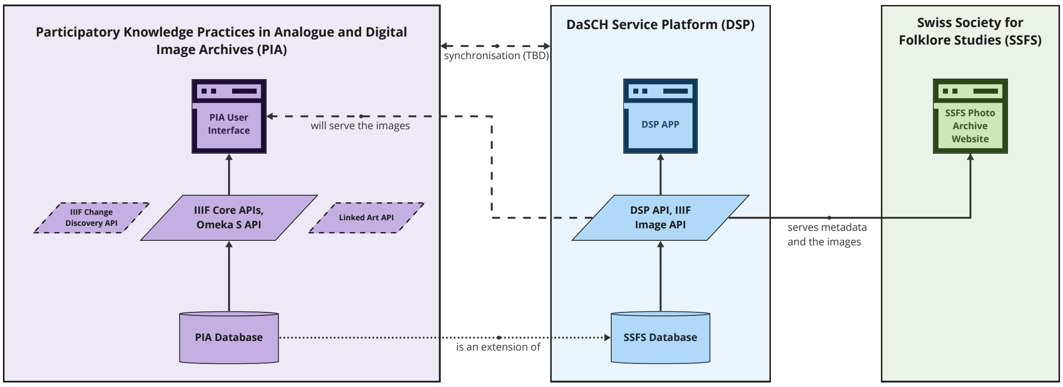

Participatory Knowledge Practices in Analogue and Digital Image Archives (PIA)
PIA User Interface
will serve the images
IIIF Change Discovery API
IIIF Core APIs, Omeka S API
Linked Art API
PIA Database
synchronisation (TBD)
is an extension of
DaSCH Service Platform (DSP)
DSP APP
DSP API, IIIF Image API
SSFS Database
serves metadata and the images
Swiss Society for Folklore Studies (SSFS)
SSFS Photo Archive Website

Figure 5: High-level overview of the PIA Infrastructure and its connection to the DaSCH Service Platform (DSP) and the official SSFS Photo Archive Website.

As can be seen in Figure 5, the PIA infrastructure runs parallel to the SSFS photo archive website, but in-between there is the DaSCH Service Platform (DSP)[10], which manages the long-term data and metadata of the SSFS and provides its own APIs (IIIF Image API and DSP API); this is where the ground truth of the digital data lies. Our database is both a selection, as it is made up of only three collections from the archives, and an extension, due to the participatory needs of the platform, of the DaSCH database. In terms of APIs, we have currently deployed the core IIIF APIs (Image and Presentation APIs) as well as a bespoke Omeka S API.

To create IIIF resources that can be displayed and compared into any IIIF-compliant viewers, we have a specific workflow based on Laravel (see Figure 6), an open-source PHP framework, which generates IIIF Presentation API 3.0 resources (`Manifests` and `Collections`) as well as associated annotations (a series of `AnnotationPage`) that comply with the Web Annotation Data Model[11]. Our main database is managed by Omeka S, which provides its own JSON-LD API. The data we collect through this API is combined with the machine learning Object Detection from vitrivr[12], which is hosted in a separate SQLite database. Next, we populate our temporary metadata collection with image-specific information provided by SIPI[13], our image server, such as pixel dimensions. The application then translates this data construct into the IIIF-specific format and serves the generated file to the client.

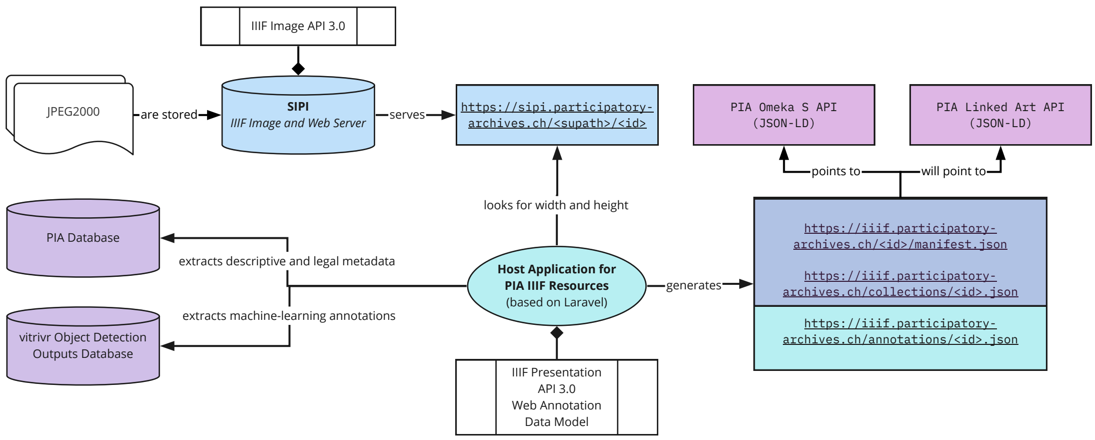

IIIF Image API 3.0
JPEG2000
are stored
SIPI
IIIF Image and Web Server
serves
https://sipi.participatory-archives.ch/<supath>/<id>
PIA Omeka S API (JSON-LD)
PIA Linked Art API (JSON-LD)
points to
will point to
looks for width and height
PIA Database
extracts descriptive and legal metadata
extracts machine-learning annotations
vitrivr Object Detection Outputs Database
Host Application for PIA IIIF Resources (based on Laravel)
generates
https://iiif.participatory-archives.ch/<id>/manifest.json
https://iiif.participatory-archives.ch/collections/<id>.json
https://iiif.participatory-archives.ch/annotations/<id>.json
IIIF Presentation API 3.0 Web Annotation Data Model

Figure 6: Overview of the PIA IIIF Resources and Annotation Workflow. To create IIIF resources that can be displayed and compared into any IIIF-compliant viewers, we have a specific host application based on Laravel, an open-source PHP framework.

As IIIF takes an agnostic approach to the provision of descriptive metadata, and has no particular intention of imparting semantics to the content it carries [28], we have decided to rely on the `seeAlso` property to link to structured metadata. Currently, our IIIF Manifests point to the Omeka S API, which is loosely based on the SSFS data model and relies heavily on

[10] https://docs.dasch.swiss/

[11] https://codeberg.org/PIA/pia-iiif-manifest-host

[12] https://vitrivr.org/

[13] https://sipi.io/

the Schema.org vocabulary. It is also planned that the IIIF Manifests will point to a forthcoming Linked Art API, which is currently under preparation[14].

The purpose of offering as many mutually agreed and shared APIs as possible from dedicated communities is to enable reuse of the data beyond the scope of PIA as well as to ensure improved data stewardship. Therefore, participation is not limited to the GUI, and we strive to provide thorough API documentation and working examples[15].

7. Future Work

In the near future the GUI will be further developed and extended. The next steps particularly focus on the elaboration of the different visual access points, such as *map*, *timeline* and *graph*. Each of these approaches brings its own possibilities and limitations that are to be tested. The use of the map, for example, to explore the photographic archives and display cartographic material is of special importance to the project because one of its three collections, *Atlas of Swiss Folklore*, has genuine geographical qualities.

We are also planning two workshops at the *Bern Academy of Arts* to test the platform's UX with different target groups. In the first workshop we will discuss with graphic and interface designers the participation functions and the visual appearance of the platform. The focus will be on how to motivate a broad audience to participate. In the second workshop, we will develop themes for future "Calls for images" with art education students who will later work, for example, in a museum or a high school.

Further action is needed in the development and integration of the metadata API and its documentation into the GUI as a user-friendly access point for programmatic interaction with the photographic archives.

For the evaluation of our APIs, and in particular for LOUD specifications, we still need to define a list of evaluation criteria. We believe that the necessary and important factors of the study are those of *learnability*, *efficiency*, *understandability*, *effectiveness*, *satisfaction*, and *readability* as there are factors commonly used to gauge an API's usability. [29]. This evaluation is expected to be carried out as part of one of the PhD theses associated with the PIA project and will take into account the specificity of the various LOUD APIs.

As the interface and its technical infrastructures matures the building of communities for citizen scientists needs to be assessed from a communicative perspective and in collaboration with the SSFS to enable sustainable usage of the collections beyond the project duration.

Last, but not least, synchronisation aspects between PIA and DaSCH will need to be clarified. It is quite possible that the data co-exist in parallel with each other while implementing a data curation and reconciliation strategy with all stakeholders, which could last after the project. Although this effort will not be straightforward, it will be streamlined by having interoperable and open systems.

[14] https://github.com/Participatory-Image-Archives/linkedart/

[15] For instance, on the Observable HQ platform: https://observablehq.com/@participatory-archives

8. Conclusion

Within PIA, we develop an environment enabling a digital workflow that starts at the original printed source and ends where experts and citizens enrich the data with their knowledge. The close dialogue between humanities researchers, archivists, and experts in design and software development, ensures a highly applicable solution upon which to engage in constructive criticism. By the transfer of archival methodologies and processes from the analogue to the digital domain, we create a sustainable aura for stored data. The innovative GUI and the integration of APIs encourage collaboration with the public and, thus, a variety of open-ended interpretive perspectives.

The participatory design approach taking place at different stages of the process – from early development to final evaluation – allows for the inclusion of a variety of perspectives in the use and reception of the SSFS archives. We hope that by giving many different groups a voice in the curation of the collections, we can contribute to a democratisation of the decision-making power of archives [30].

As we show, the sustainability of data and digital tools is closely related to the application; we go beyond open data by demonstrating the power of standardised APIs, namely those that adhere to the LOUD design principles. The possibility to enrich data makes data sustainable and increases the attraction of digital infrastructures.

Acknowledgments

We would like to express our sincere gratitude to all the colleagues, partners, and individuals involved in the PIA research project, whose invaluable contributions, support, and collaboration were instrumental in the successful completion of this paper.

Participatory Knowledge Practices in Analogue and Digital Image Archives (PIA) is supported by the Swiss National Science Foundation (SNSF) as part of the Sinergia funding scheme (contract no. CRSII5_193788): https://data.snf.ch/grants/grant/193788.

References

[1] T. Kärberg, K. Saarevet, Transforming User Knowledge into Archival Knowledge, D-Lib Magazine 22 (2016). URL: http://www.dlib.org/dlib/march16/karberg/03karberg.html. doi:10.1045/march2016-karberg.

[2] G. Hinchcliffe, M. Whitelaw, The Corley Explorer, State Library of Queensland (2018). URL: https://openresearch-repository.anu.edu.au/handle/1885/206061. doi:10.25911/5f0c3898c50a8.

[3] N. Graf, Georeferenzierung in sMapshot, ETH Zürich, ETH-Bibliothek, Bildarchiv, Bern, Switzerland, 2022, p. 10. URL: https://www.research-collection.ethz.ch/handle/20.500.11850/554517. doi:10.3929/ethz-b-000554517.

[4] M. Baggett, R. Gibbs, Historypin and Pinterest for Digital Collections: Measuring the Impact of Image-Based Social Tools on Discovery and Access, Journal of Library Administration 54 (2014) 11–22. doi:10.1080/01930826.2014.893111.

[5] A. Carron, Le crowdsourcing pour enrichir une plateforme d'archives participatives: notreHistoire.ch, Travail de bachelor, HES-SO University of Applied Sciences and Arts, Haute école de gestion de Genève, Carouge, Switzerland, 2018. URL: https://sonar.ch/global/documents/314759.

[6] M. Ridge, Making digital history: The impact of digitality on public participation and scholarly practices in historical research, phd, The Open University, 2016. URL: http://oro.open.ac.uk/45519/.

[7] F. Schmoll, Richard Weiss : Skizzen zum internationalen Wirken des Schweizer Volkskundlers, Schweizerisches Archiv für Volkskunde/ Archives suisses des traditions populaires 2009 (2009) 15–32. URL: https://www.e-periodica.ch/digbib/view?pid=sav-001:2009:105::26. doi:10.5169/SEALS-118266.

[8] P. Pfrunder, Ernst Brunner: Photographien, 1937-1962, 2. aufl ed., Schweizerische Gesellschaft für Volkskunde ; Offizin, Basel : Zürich, 1995.

[9] R. Sanderson, Shout it Out: LOUD, 2018. URL: https://www.slideshare.net/Europeana/shout-it-out-loud-by-rob-sanderson-europeanatech-conference-2018.

[10] J. A. Raemy, Améliorer la valorisation des données du patrimoine culturel grâce au Linked Open Usable Data (LOUD), in: N. Lasolle, O. Bruneau, J. Lieber (Eds.), Actes des journées humanités numériques et Web sémantique, Les Archives Henri-Poincaré - Philosophie et Recherches sur les Sciences et les Technologies (AHP-PReST); Laboratoire lorrain de recherche en informatique et ses applications (LORIA), Nancy, France, 2022, pp. 132–149. doi:10.5451/unibas-ep89725.

[11] S. Snydman, R. Sanderson, T. Cramer, The International Image Interoperability Framework (IIIF): A community & technology approach for web-based images, in: Archiving Conference, volume 2015, IS&T, Los Angeles, CA, 2015, pp. 16–21. URL: https://purl.stanford.edu/df650pk4327.

[12] M. D. Wilkinson, M. Dumontier, I. J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J.-W. Boiten, L. B. da Silva Santos, P. E. Bourne, J. Bouwman, A. J. Brookes, T. Clark, M. Crosas, I. Dillo, O. Dumon, S. Edmunds, C. T. Evelo, R. Finkers, A. Gonzalez-Beltran, A. J. Gray, P. Groth, C. Goble, J. S. Grethe, J. Heringa, P. A. 't Hoen, R. Hooft, T. Kuhn, R. Kok, J. Kok, S. J. Lusher, M. E. Martone, A. Mons, A. L. Packer, B. Persson, P. Rocca-Serra, M. Roos, R. van Schaik, S.-A. Sansone, E. Schultes, T. Sengstag, T. Slater, G. Strawn, M. A. Swertz, M. Thompson, J. van der Lei, E. van Mulligen, J. Velterop, A. Waagmeester, P. Wittenburg, K. Wolstencroft, J. Zhao, B. Mons, The FAIR Guiding Principles for scientific data management and stewardship, Scientific Data 3 (2016) 160018. URL: https://www.nature.com/articles/sdata201618. doi:10.1038/sdata.2016.18.

[13] R. Sanderson, Cultural Heritage Research Data Ecosystem, 2020. URL: https://www.slideshare.net/azaroth42/sanderson-cni-2020-keynote-cultural-heritage-research-data-ecosystem.

[14] M. Appleby, T. Crane, R. Sanderson, J. Stroop, S. Warner, IIIF Design Principles, 2018. URL: https://iiif.io/api/annex/notes/design_principles/.

[15] R. Sanderson, P. Ciccarese, H. Van de Sompel, Designing the W3C open annotation data model, in: Proceedings of the 5th Annual ACM Web Science Conference, WebSci '13, Association for Computing Machinery, New York, NY, USA, 2013, pp. 366–375. URL: https://doi.org/10.1145/2464464.2464474. doi:10.1145/2464464.2464474.

[16] M. Doerr, The CIDOC Conceptual Reference Module: An Ontological Approach to Semantic Interoperability of Metadata, AI Magazine 24 (2003) 75–75. URL: https://ojs.aaai.org/index.php/aimagazine/article/view/1720. doi:10.1609/aimag.v24i3.1720, number: 3.

[17] D. Newbury, LOUD: Linked Open Usable Data and linked.art, in: 2018 CIDOC Conference, International Council of Museums, Heraklion, Greece, 2018, pp. 1–11. URL: https://cidoc.mini.icom.museum/wp-content/uploads/sites/6/2021/03/CIDOC2018_paper_153.pdf.

[18] I. Jacobs, N. Walsh, Architecture of the World Wide Web, Volume One, 2004. URL: https://www.w3.org/TR/webarch/.

[19] M. Appleby, T. Crane, R. Sanderson, J. Stroop, S. Warner, IIIF Change Discovery API 1.0.0, 2021. URL: https://iiif.io/api/discovery/1.0/.

[20] M. Ridge, There's a new viewer for digitised items in the British Library's collections, 2016. URL: https://blogs.bl.uk/digital-scholarship/2016/12/new-viewer-digitised-collections-british-library.html.

[21] J. A. Raemy, The International Image Interoperability Framework (IIIF): raising awareness of the user benefits for scholarly editions, Bachelor's thesis, Haute école de gestion de Genève, Geneva, Switzerland, 2017. URL: https://sonar.ch/hesso/documents/314853.

[22] J. Bergold, Partizipative Forschung und Forschungsstrategien, 2013. URL: https://www.buergergesellschaft.de/fileadmin/pdf/gastbeitrag_bergold_130510.pdf.

[23] D. A. Norman, Human-centered design considered harmful, Interactions 12 (2005) 14–19. URL: https://dl.acm.org/doi/10.1145/1070960.1070976. doi:10.1145/1070960.1070976.

[24] A. Cooper, R. Reimann, D. Cronin, A. Cooper, About face 3: the essentials of interaction design, [3rd ed.], completely rev. & updated ed., Wiley Pub, Indianapolis, IN, 2007.

[25] A. Poikola, K. Kuikkaniemi, H. Honko, MyData – A Nordic Model for human-centered personal data management and processing, Finnish Ministry of Transport and Communications, Helsinki, Finland, 2015. URL: http://urn.fi/URN:ISBN:978-952-243-455-5.

[26] M. Ridge (Ed.), Crowdsourcing our cultural heritage, Digital research in the arts and humanities, first issued in paperback ed., Routledge, London New York, 2017.

[27] B. Frost, Atomic design, Brad Frost, Pittsburgh, Pennsylvania, 2016.

[28] J. A. Raemy, A. Demleitner, Implementation of the IIIF Presentation API 3.0 based on Software Support: Use Case of an Incremental IIIF Deployment within a Citizen Science Project, in: Digital Heritage. Progress in Cultural Heritage: Documentation, Preservation, and Protection, Lecture Notes in Computer Science, Springer International Publishing, Cham, 2023. This paper is part of the EuroMed2022 Proceeding which will be published in 2023.

[29] I. Rauf, E. Troubitsyna, I. Porres, A systematic mapping study of API usability evaluation methods, Computer Science Review 33 (2019) 49–68. URL: https://www.sciencedirect.com/science/article/pii/S1574013718301515. doi:10.1016/j.cosrev.2019.05.001.

[30] C. Franzoni, H. Sauermann, Crowd science: The organization of scientific research in open collaborative projects, Research Policy 43 (2014) 1–20. URL: https://linkinghub.elsevier.com/retrieve/pii/S0048733313001212. doi:10.1016/j.respol.2013.07.005.

Untapped data resources

Applying NER for historical archival records of state authorities

Venla Poso[1], Tanja Välisalo[1], Ida Toivanen[1], Antero Holmila[1] and Jari Ojala[1]

[1]*University of Jyväskylä, Seminaarinkatu 15, Jyväskylä, Finland*

Abstract

Archives around the world are digitising their material at a growing speed. The National Archives of Finland launched a mass digitisation process in 2019 aiming to digitise vast amounts of state authority archives. In order to improve the access and use of this data by researchers, we present the data transfer process of state authority data and the development of named entity recognition (NER) for enriching and using archival data from state authorities. In this process, we have developed two new named entities that are not included in published NER models for the Finnish language. This work is conducted as part of the DARIAH-FI infrastructure.

Keywords

named entity recognition, archival records, state authority archives, tool development

1. Introduction

Archives around the world are digitising their material at a vastly growing speed. This means that massive amounts of records will be made available to researchers in various fields of study. This opens up a wide range of possibilities for researchers. For historical research in particular, this kind of mass digitisation is important in helping prevent the risk of 'source myopia', which can result from very limited types of data being available in digital format[1].

The National Archives of Finland launched a mass digitisation project in 2019 ultimately aiming to digitise 135 shelf kilometres of state authority records with the intent of destroying the original documents [2]. The mass digitisation project includes various areas of development such as improving the quality of optical character recognition (OCR) and segmentation detection [3]. The aim of mass digitisation is not only to make the archives more accessible for state authorities but also to advance the possibilities of using archival material in various fields of research. Similar large-scale digitisation has been underway or forthcoming in several other national archives as well, such as National Archives of the Netherlands, State Archives of Belgium, The Swedish National Archives, and US National Archives and Records Administration (e.g., [4]), which makes developing the usability of archival data from the research perspective of particular importance.

DHNB2023 | Sustainability: Environment - Community - Data. The 7th Digital Humanities in the Nordic and Baltic Countries Conference. Oslo – Stavanger – Bergen, Norway. March 8–10, 2023.

venla.s.m.poso@jyu.fi (V. Poso); tanja.valisalo@jyu.fi (T. Välisalo); ida.m.toivanen@jyu.fi (I. Toivanen); antero.holmila@jyu.fi (A. Holmila); jari.ojala@jyu.fi (J. Ojala)

0000-0001-8678-4683 (T. Välisalo); 0000-0003-2456-7223 (A. Holmila); 0000-0002-4348-8857 (J. Ojala)

DHNB Publications, DHNB2023 Conference Proceedings, https://journals.uio.no/dhnbpub/issue/view/875

Advancing the use of archival data for research is the objective of the FIN-CLARIAH[1] infrastructure project, which focuses on finding the best practices for enriching and accessing the recently digitised data in the National Archives of Finland. Making the digitised archives more accessible and usable for research purposes demands enriching the data in various ways. One way to make the data more usable for researchers is to utilise a natural language processing (NLP) task called Named Entity Recognition (NER). NER is an information extraction method, which is used to identify different types of entities, such as persons, organisations, places, dates, times, or events, from unstructured text. The development of NER for state authority archives in this project will lead to the deployment of NER models as open-source tools for researchers, as well as integrated in the services of the National Archives of Finland.

Before starting our project, some of the digitised data could be accessed through the National Archives' online service ASTIA[2]. ASTIA is meant for browsing and accessing digitised documents through a web browser. ASTIA interface consists of a split screen with the original document image on the left and the text file on the right. Individual documents can be downloaded as 1) JPG, 2) PDF, and/or 3) ALTO XML files, and larger document collections as separate JPG and ALTO XML files. These options are certainly diverse and sufficient enough for most current needs with openly available documents, especially those with traditional qualitative research approaches. However, for many digital humanities and social sciences methods (SSDH), particularly those identified as big data methods, this is not the most convenient or efficient technique for accessing the data. Additionally, sensitive data needs a secure method for accessing and browsing it. Therefore, there is a need for other forms of data transfer.

In this paper, we will describe the design of the NER process for digitised state authority archives and consider the potential benefits and challenges of using NER in historical research with this type of archival data. As part of the design process, we created a survey directed at researchers within the fields of humanities and social sciences in order to bring new perspectives to the possibilities of using NER. We will also report the results of this survey and how they were incorporated in the process.

2. Digital history

Digital humanities (DH) is an interdisciplinary field where computational studies and humanities meet [5]. It is debatable if fields such as digital history, corpus linguistics and other digitally oriented research factions fall under DH or whether they are co-managing the field. Nevertheless, in this paper we concentrate on the field of digital history. Similar to the definition of DH, digital history is a diverse field which can be determined from the perspective of the subject of study or the methodological approach [6, 7]. Hannu Salmi [6] has defined digital history as "an approach to examining and representing the past; it uses new communication technologies and media applications and experiments with computational methods for the analysis, production

[1] FIN-CLARIAH infrastructure project 2022-2023, funded by the Academy of Finland, comprises FIN-CLARIN and DARIAH-FI, which are part of European research infrastructures CLARIN ERIC and DARIAH-EU. FIN-CLARIAH aims to develop processes, methods and tools for processing unstructured text in social sciences and humanities research.

[2] https://astia.narc.fi/uusiastia/

and dissemination of historical knowledge." The ambiguousness of the field has been captured by Seefeldt and Thomas[7]:

> On one level, digital history is an open arena of scholarly production and communication, encompassing the development of new course materials and scholarly data collections. On another, it is a methodological approach framed by the hypertextual power of these technologies to make, define, query, and annotate associations in the human record of the past.

We emphasise the methodological approach and centre our attention to the development and possibilities of the computer-assisted methods and tools. Digitised data holds a multitude of possibilities for historians, in addition to remote access in itself [8]. Within the field of digital history researchers have found new angles to old subjects. The multitude of methodological paths to choose from is forever expanding. Although the digital history approaches might offer some objectivity to the process and reveal unseen patterns from the used data, the researchers' choices are still on the lead. Within topic modelling, sentiment analysis, text network analysis or other data mining options there are various choices to make which affect the outcome [9]. Using a computer-assisted approach helps the researcher to avoid 'cherry picking', which means that researcher finds parts of the data to support their preliminary hypothesis and disregard other viewpoints [10, 11]. Also, the major promise for digital history in the historian's perspective is the possibility of examining vastly more data than has been the typical practice among historians. As the amount of data has exploded over the last couple of decades, it is evident that in the future historians will need a totally new toolset for the practice of their discipline. According to Thaller[12]:

> The Humanities have since their earliest inception always been focusing on the ability to draw a maximum of conclusions from a rather limited amount of information they could access physically. The [sic] only start to notice that this barrier has broken down. The primary qualification of a Humanities' researcher of the year 2050 will not be, how to lovingly extract insights from a few isolated bits of information, but how to meaningfully integrate the information contained in the largest possible set of data.

However, despite the promise of DH, digital history is not a short-cut for new insights and ideas. As David Blei [13] has written:

> . . . statistical models are meant to help interpret and understand texts; it is still the scholar's job to do the actual interpreting and understanding [. . .] the hope is that the model helps point us to [. . .] evidence. Using humanist texts to do humanist scholarship is the job of a humanist.

The field of digital history is not without its problems. Researchers must understand the assisting technologies and what the information extracted with them actually tells. In addition to that, we are aware of the possible pitfalls, such as the "virtual dismemberment" [8] of archives, when single documents can become separate from the collection they are part of, and thus, lose some of their interpretative potential. As Lara Putnam has reminded, for the first time, historians (and other humanists), can find vast amounts of data without understanding the genealogy of the data - that is, who has created it and why; what is its place in the larger hierarchy of the archive and so on: "Web-based full-text search decouples data from place [. . .] for the first time, historians can find without knowing where to look.[14]"

Part of conquering these difficulties with digital history is choosing the right computer-assisted methods and understanding the limits and possibilities that they offer. As the DH field is booming and new methods, analysis tools and datasets are mushrooming, the challenge for newcomers is simply where to start. In this paper, we concentrate on the NER and its applications in the field of digital history. Locating the different entities from the archival material offers valuable information for the researchers in various stages of the research process, and enriching the archival data with NER can be useful to the archive itself.

3. Named Entity Recognition

Named entity recognition (NER) was originally developed as a form of information extraction (IE)[15]. Current use of NER exceeds the original purpose, and it has been utilised in a wide range of different NLP tasks. The core task of NER has remained the same, as locating and naming predefined entities[16], but the development of new applications continues to be a popular field in natural language processing (e.g., [17, 18, 19, 20, 21, 22, 23, 24]).

Nordic and Baltic languages have quite sufficient NER models and corpora (see [25, 26, 27]). For Finnish NER, there are three notable corpora: 1) The FiNER[24], 2) Turku NER[28], and 3) TurkuONE[29]. The FiNER corpus is mainly based on single-domain text, technology news from the magazine Digitoday, so its contribution is limited when moving on to a different domain. In FiNER, there are six entity groups (organisation, location, person, product, event, date). The Turku NER corpus took this into account, by being constructed from various domains and text types. The Turku NLP group has created TurkuONE, a new fine-grained NER corpus, which combines and extends the two previously published corpora. The most notable difference is in the used NER categories, which have been revised to match international standards. This means that the number of categories has been changed from six to eighteen different entity groups (based on OntoNotes 5.0, see [30]). It is important to note that the definitions for entity groups differ from the older NER corpora to TurkuONE, since the fine-grained version divides some of the categories, such as location, to smaller sections, such as facility, geopolitical location, and other locations.

The aforementioned NER corpora have been used in the training and testing of different NER models. For example, the first two have been used to train and test FiNER tagger, which is a rule-based NER model[31, 24]. As seen in Table 1, the FiNER tagger performed reasonably well for FiNER corpus, but when tested with Turku NER corpus, the F1 score, which is used to measure the accuracy of machine learning models, dropped remarkably. There are also differences in performance of FiNER tagger when it comes to different entity groups. For example, when FiNER tagger was tested with a Wikipedia test set, the overall F1 score dropped to 79.91, while the scores for PRO, ORG and EVENT classes were close to or under 60[24].

Recently the FiNER tagger has been outperformed by other models. Development of the BERT model[32] has brought new possibilities to the NLP field. BERT can be used as a backbone in

Table 1
FiNER tagger scores, tested with Turku NER corpus[28] and FiNER corpus/Digitoday test set[24]

NER corpus	F1-score	Prec.	Rec.
Turku NER corpus	74.08	77.16	71.24
FiNER corpus	85.20	90.79	80.25

Table 2
FinBERT scores, tested with TurkuONE corpus and Turku NER corpus[28, 29].

NER corpus	F1-score	Prec.	Rec.
TurkuONE corpus	92.99	92.58	93.41
Turku NER corpus	91.65	90.87	92.44

tasks such as NER. The Finnish version of BERT, the FinBERT model, is pre-trained from scratch on Finnish data[33]. FinBERT has been used in testing the most recent NER corpora[28, 29] and performed well in both cases. As seen in Table 2, FinBERT seems to perform better on the test set from the TurkuONE corpus, implicating that the difference in performance is explained by the extended number of entity groups. Drawing from these results, we hope to build a NER tool for state authority archives using FinBERT.

4. Applications of NER in historical research

“...historian exhausting the records before they exhaust the historian.[34]"

Researchers can rarely control the processes that lead to the formation of the archives which they choose to examine. Independent of the type of the archive, decisions need to be made on what is preserved and in what quantity. Additionally, when it comes to the state authority and administrative processes, the amount of documentation has increased over the course of history [35]. The variation in archival practices over the years and in different institutions have added their own twist to the overabundance of archival material[36]. This has led to a situation where serendipity might play an important role in researchers’ work[36]. NER can assist in improving information retrieval and making the process of selecting the relevant archival material for research purposes a more traceable process[37]. Common search functions enable researchers to find information with a specific and controlled vocabulary, which most times gives exactly what the researcher is looking for. However, using a search based on NER, the search produces a wider set of results, which might reveal something beyond what the researcher expected. Nevertheless, the NER-based search results still need to be analysed by the researcher and all results do not automatically hold meaning for the research. Further analysis can include determining what is the meaning of each entity in a passage or document.

Improving information retrieval for archival data is and will be an important part of NER, but entity recognition offers a wide set of other possibilities as well. It can work as a basis or in connection with other digital humanities and digital history methods. As the three fundamental entity types, person, organisation and location (MUC-6 competition[38]) often are most frequently mentioned across the document types, the recognition of them also seems to be the most trustworthy and most often used as a basis for research. For example, [39] studied canonisation of cultural memory in the online audiovisual archives of the Finnish Broadcasting Company using most frequently appearing entities of names, places and events in the archival metadata. As another example, Erik Edoff used frequency of place entities in newspapers from different eras in the late 1900s to see if new technologies really made the world smaller[40]. Place entities revealed that contrary to what is generally perceived, newspapers included more mentions of places in the local region than of far-away places, which would not speak for a smaller world, but a tighter and constant connection with the neighbouring cities. Place entities can be further explored using geomapping[41], where the place entities are combined with map visualisations. In historical research, geomapping has often been used to depict temporal changes. For example, Clifford[42] studied the development of an industrial ecosystem with the aid of geomapping.

In addition to locations, the entity category "person" has been a valuable source in historical research. The frequency counts of specific persons can be used to detect their cultural meaning when the data supports this kind of hypothesis[39]. A more complex NLP task compilation where NER has been used as a basis for the analysis, is presented in a study by Fields et al.[43] on the Ottoman-Iraqi personal diaries. There, named entity recognition is used alongside network analysis to map a person's daily life, community structure, and social relations.

5. Survey on named entities

In order to gain a deeper understanding of potential needs for NER based tool development, we conducted an online survey targeted at researchers interested in using state authority archives. The aim of the survey was to support sustainable NER development, where diverse research perspectives would be taken into consideration starting from the beginning of the process. The survey was distributed to researchers in the fields of history and social sciences in particular, through universities, research societies and conferences between November 2022 and February 2023.

The survey gathered 57 responses from Finnish researchers. The respondents represented multiple research areas with an emphasis on history (see Table 3). The survey described briefly what named entity recognition is, and the respondents were given a list of named entities and asked which named entities they felt were most useful for them. The results give a general idea about the entity types that the researchers would prefer, although it is important to note that the number of survey respondents was fairly low.

The survey respondents were also asked to self-evaluate their previous experiences with digital research methods on a given scale (Table 4). Majority of respondents (91.2%) described themselves as having at least little knowledge of digital research methods. When asked to

Table 3
Participants according to their academic discipline (n=57).

Academic discipline	f	%
History	31	53.4
Other Humanities	12	20.7
Social Sciences	11	19.0
Other disciplines	2	3.4
Multiple disciplines	1	1.7
Total	57	100.0

Table 4
Previous experience with digital research methods (n=57).

Previous experience with digital research methods	N	%
None	5	8.8
Little	13	22.8
Some	27	47.3
Much	7	12.3
Very much	5	8.8
Total	57	100.0

elaborate, half of the respondents (27; 49.1%) gave a more detailed description. Most common were mentions of using digital data (e.g., digitised archival data) with 10 respondents mentioning only this form of digital research methods. Other common responses described using digital analysis tools (9 mentions) or digital search tools (8 mentions). There were also mentions of using digital databases, digital tools for collecting data (e.g., survey tools), and digital data management tools.

The main question in the survey was on the perceived usefulness of different named entities in regard to the state authority archives. The respondents were presented with 19 different entities along with a few examples, and they were asked to evaluate how useful these entities were on a 4-point scale with options 'very useful', 'useful', 'possibly useful', and 'not useful'. All respondents answered this question but not all entities. The response rate per entity differed between 89.5% and 100%. The entities considered 'very useful' were Journal number, Date, Nationality, religious or political group, Geopolitical location, Organisation, and Person (see Figure 1).

Respondents were also asked what other things or entities might be useful to recognize in the survey data. The responses (n = 18) were quite diverse and some of the things mentioned can already be solved through existing categories; for instance, different joint municipalities, when directly mentioned, would fall under the organisation (ORG) entity. Many suggestions were also so specific that they would be more easily attained by traditional search functions (e.g., nuclear power, inflation). One suggestion made by more than one respondent was profession or professional title. However, this question also yielded an interesting result pertaining to researcher needs: several respondents mentioned needs pertaining to the metadata, such as the

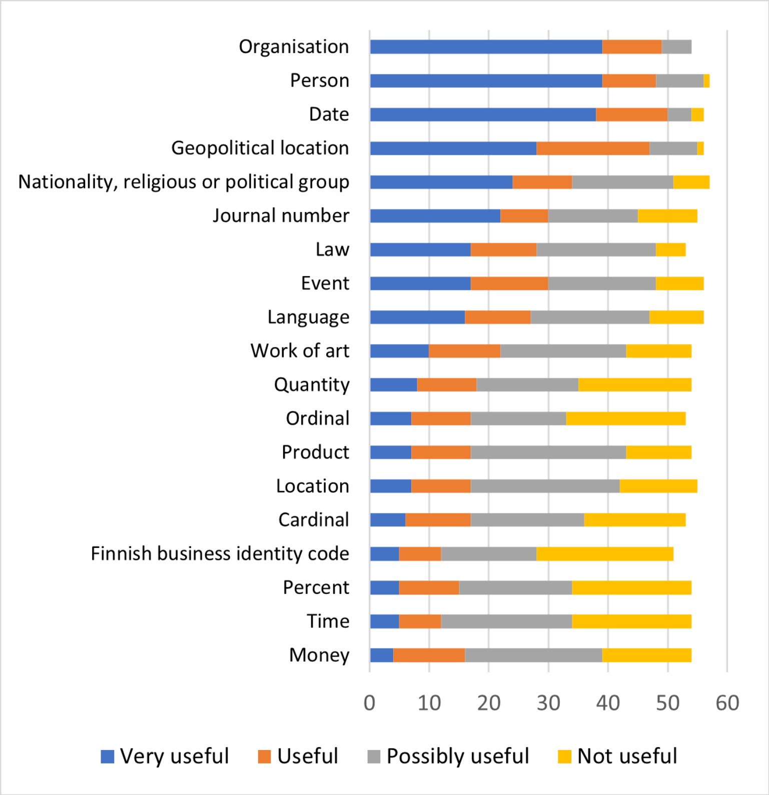


Figure 1: The perceived usefulness of different entities by survey respondents.

need to know what type of documents is in question or whether the named entity (NE) is in a heading or body text.

6. Developing NER for digitised state authority archives

In applying named entity recognition for digitised state authority data, our objective is to pave the way for researchers who wish to use archival data in novel ways. The National Archives of Finland is currently digitising masses of archival documents, which means there is a need for rapid and consistent metadata generation. Our process for named entity recognition follows similar steps as the previous work done on Finnish NER. Existing Finnish NER models, however, are not tailored to archival data, so developing NER further is crucial. We aim to document, evaluate and report the whole development process for faster deployment of mass digitised archival data in research. We approach the process from the perspective of creating tools for the end-user who would like to better utilise archival data in their research.

In the following, we will first describe the process of accessing state authority records in the National Archives, and then the annotation of named entities. We will describe the particular demands that state authority archives make on the process.

6.1. Accessing data from the National Archives

As a pilot data for our project, we used the mass digitised archives of Finland's Ministry of Economic Affairs and Employment. In order to access the digitised data, the research institution makes a data transfer agreement with all the parties involved - NARC, data owner, and CSC (see Figure 2). CSC, or IT Center for Science, is a Finnish government owned company, which provides higher education and research institutions research infrastructure services such as supercomputers and servers. As this project was aimed at research infrastructure development, particularly developing methodological tools, it was not a typical research project where agreement and licensing processes between state authorities and archives are already established. Additionally, the large quantities of data made it necessary to use third party services for data transfer, which was another departure from the more common forms of cooperation between these organisations. These novel features of this project introduced new demands also to the formal agreements and licences.

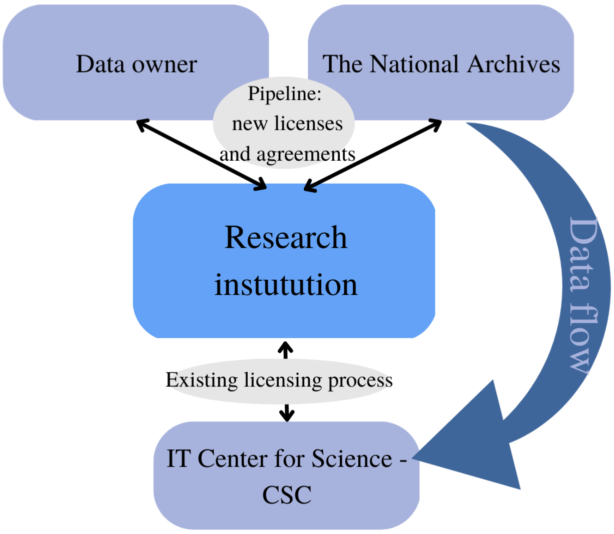

Data owner
The National Archives
Pipeline: new licenses and agreements
Research instutution
Data flow
Existing licensing process
IT Center for Science - CSC

Figure 2: The agreements and data flow between different actors in making the mass digitised state authority archives available to researchers.

The pilot data contains data from 1999-2017 that consists mostly of typewritten materials. The data includes documents in several languages, but we focus only on data in Finnish. The pilot data entails personal details and confidential data, which places particular demands on the agreements and the security of transferring and processing data. The pilot data is transformed from image of text into machine-readable text using OCR technology. The quality of digitised material often varies regarding the OCR, metadata and data structure[44]. The document type can also heavily influence the results yielded by applying NER. For example, documents containing an abundance of tables have been found to affect NER results with archival data (e.g., [41]).

Currently, the quality of the OCR is generally fairly sufficient for research purposes. Particular challenges of applying OCR for the pilot data at this stage include hand-written data, non-conventional document layout, headings, and special characters. Handwritten text recognition (HTR) is not yet included in the mass digitisation process. Non-conventional document layout, for example, a 2-column layout, is not recognized in the OCR process as separate text areas, which results in the output text rows consisting of fragments from both columns, breaking the original sentences and paragraphs. Headings in the text are not consistently recognized as such, but are sometimes combined with the adjacent paragraph sign, and are also inconsistently recognized as text. In addition to these difficulties, similar to the OCR errors reported in the literature[45], in the pilot data there also may be misrecognized, missing or extra characters, as well as nearby words grouped together, or division of a word into several subwords. In previous studies, it is suggested that OCR quality should be adequate in order to develop state-of-the-art NER[46]. In this context, the data having OCR generated errors makes it noisy - that is, the data is at least partially corrupted. The difficulties with noisy data may seem impenetrable, but as Fridlund et al.[1] (2020) have argued: "If we limited our research to clean datasets, very little would be accomplished." Part of this process is to document the effects of OCR issues on NER in this particular data type.

6.2. Annotation process

Based on our NER survey, researchers found Journal number, Date, Nationality, religious or political group, Geopolitical location, Organisation, and Person as the most useful named entities. NARC executed a similar survey directed to seven different authorities in which Journal number and Finnish business identity code were considered 'very useful'. Building from both survey results, we ended up with ten entity categories: person (PER), organisation (ORG), location (LOC), geopolitical location (GPE), product (PRO), event (EVENT), date (DATE), journal number (JON), Finnish business identity code (FIBC), and nationality, religious and political group (NOPR). Journal number (JON) and Finnish business identity code (FIBC) are new entities created and defined specifically in this project. In addition to the survey results, we based our entity categories on previous work on Finnish NER[24, 28, 29].

Data annotation is necessary to train deep learning models for the NER task. There are several steps in the annotation process. The annotation scheme we use is BIO/IOB2 format. First, we preprocess the data. This includes changing the format from AltoXML to CSV, as well as filtering out the data that is in another language than Finnish. Then, we add pseudo-labels for entities using a previously trained NER model by TurkuNLP. After pseudo-labelling, there are still two entity groups (i.e., journal number and Finnish business identity code) that are not pre-annotated. After the first round of model-aided annotations, we manually examine the data and pseudo-labels and make necessary additions and corrections. For the whole annotation process, CSC's computational services are used in order to handle the data in a secure environment. CSC offers computational services for sensitive data (e.g., SD Desktop and SD Connect), which makes it possible to develop the methodological tools needed.

While some researchers have measured inter-annotator agreement when using manual annotation as part of the NER process (e.g., [47]), we opted for creating precise NER annotation guidelines before the annotation began and refining the guidelines through close communication between annotators as the annotation was taking place. Similar decision was made when making the TurkuONE corpus, where only one annotator was involved in the process[29]).

After the manual annotation phase, the annotated data is used for model training. The annotation tool is then tested with other data in cooperation with researchers to define the need for improvements in the process. Re-training and testing are executed if needed. The final aim is to distribute the NER tool for researchers to use.

7. Conclusions and discussion

Mass digitisation produces previously unseen quantities of archival data in a uniform digital format. However, the methods and tools for using this data are still under development. In this paper, we have reported the two main contributions of our work thus far. First, we have defined the process of accessing and using the state authority records from the National Archives in order to make the process available as a benchmark for future researchers. Second, we have applied new named entities to the NER annotation process for Finnish language text. In our exploration of using NER for historical research and state authority data in particular, we started by mapping the existing needs of researchers within the field. Based on the survey and the survey by NARC, we identified the need for two named entities, Journal number (JON) and Finnish business identity code (FIBC), which were not included in existing NER models.

State authority archival records are often arranged based on the types of content (letters, minutes etc.), rather than by the topics or themes present inside the documents. This makes named entity recognition particularly useful for researchers as it can help recognize the documents that are useful with a particular research topic. NER provides a multitude of possibilities for researchers. For example, it can help identify different actors (e.g., advocacy groups) affecting different processes. NER also enables tracing policy trends and effects of local/world events in different processes. It can also help identify regional variations, which can further be examined using visualisations. For example, researchers could explore whether certain areas are emphasised when implementing certain policies or distributing funds, or which particular foreign countries or cities are present as points of comparison or partners in certain areas. Furthermore, NER can open up new perspectives on the state authority practices when used for open exploration of archive contents.

Next steps in the process include testing the developed NER model with a wide variety of state authority data as well as other types of archival data. Applications of the NER model should include combining NER results with metadata extracted using other techniques, such as identifying document structures. As state authority data also includes data and documents in other languages, especially Swedish, a multilingual NER development is one possible direction for further tool development. Future work on named entity recognition and analysis tools based on NER should entail actively utilising new technological advancements emerging in the fields of natural language processing and machine learning.

References

[1] M. O. Fridlund, Mats, P. Paju, Digital Histories: Emergent Approaches within the New Digital History, Helsinki University Press, 2020. URL: https://doi.org/10.2307/j.ctv1c9hpt8.

[2] T. Hölttä, V. Kajanne, No more new archive buildings – mass digitisation and retroactive digitisation improve the accessibility of material, in: J. Nuorteva, P. Happonen (Eds.), The National Archives of Finland Strategy 2025: Perspectives for the future, The National Archives of Finland, 2020, pp. 14–15. URL: "https://kansallisarkisto.fi/documents/141232930/153230445/KA_Strategy_2025_eng.pdf".

[3] T. N. A. of Finland, Dalai - using artificial intelligence to improve the quality and usability of digital records, ???? URL: https://kansallisarkisto.fi/en/dalai-en.

[4] L. Hirvonen, Survey of Digitization in Archives, The National Archives of Finland, 2017. URL: https://kansallisarkisto.fi/documents/141232930/150411434/Liite_2_Digitization_Survey_2017.pdf.

[5] T. Schwandt (Ed.), Digital Methods in the Humanities: Challenges, Ideas, Perspectives, Bielefeld University Press, 2021.

[6] H. Salmi, What is Digital History?, Polity Press, Cambridge, UK, 2021.

[7] D. Seefeldt, W. G. Thomas, What is Digital History? A Look at Some Exemplar Projects, Perspectives on History (2009). URL: https://www.historians.org/research-and-publications/perspectives-on-history/may-2009/what-is-digital-history.

[8] B. Ogilvie, Scientific Archives in the Age of Digitization, Isis 107 (2016) 77–85. URL: https://www.journals.uchicago.edu/doi/full/10.1086/686075.

[9] S. Ramsay, Databases, in: S. Schreibman, R. Siemens, J. Unsworth (Eds.), A Companion to Digital Humanities, Blackwell Publishing Ltd, 2004, pp. 177–197. doi:10.1002/9780470999875.ch15.

[10] P. Baker, C. Gabrielatos, M. KhosraviNik, M. Krzyżanowski, T. McEnery, R. Wodak, A useful methodological synergy? combining critical discourse analysis and corpus linguistics to examine discourses of refugees and asylum seekers in the uk press, Discourse & Society 19 (2008) 273–306. doi:10.1177/0957926508088962.

[11] V. Koller, G. Mautner, Computer applications in critical discourse analysis, in: C. Coffin, A. Hewings, K. O'Halloran (Eds.), Applying English grammar: functional and corpus approaches, Routledge, London, 2020, pp. 216–228.

[12] M. Thaller, The humanities are about research, first and foremost; their interaction with computer science should be too, in: C. Biemann, G. R. Crane, C. D. Fellbaum, A. Mehler (Eds.), Computational Humanities - bridging the gap between Computer Science and Digital Humanities (Dagstuhl Seminar 14301), volume 4 of *Dagstuhl Reports*, Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, Dagstuhl, Germany, 2014, pp. 80–111. doi:10.4230/DagRep.4.7.80.

[13] D. M. Blei, Topic modeling and digital humanities, Journal of Digital Humanities 2 (2012). URL: https://journalofdigitalhumanities.org/2-1/topic-modeling-and-digital-humanities-by-david-m-blei/.

[14] L. Putnam, The Transnational and the Text-Searchable: Digitized Sources and the Shadows They Cast, The American Historical Review 121 (2016) 377–402. doi:10.1093/ahr/121.2.377.

[15] D. D. Palmer, D. S. Day, A statistical profile of the named entity task, in: Fifth Conference on Applied Natural Language Processing, Association for Computational Linguistics, Washington, DC, USA, 1997, pp. 190–193. doi:10.3115/974557.974585.

[16] A. V. K. S. A. O.-B. Marco Humbel, Julianne Nyhan, The effect of morphology in named

entity recognition with sequence tagging, Journal of Documentation 77 (2021) 1223–1247. doi:10.1108/JD-02-2021-0032.

[17] G. Lample, M. Ballesteros, S. Subramanian, K. Kawakami, C. Dyer, Neural architectures for named entity recognition, in: Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, San Diego, California, 2016, pp. 260–270. URL: https://aclanthology.org/N16-1030. doi:10.18653/v1/N16-1030.

[18] X. Ma, E. Hovy, End-to-end sequence labeling via bi-directional LSTM-CNNs-CRF, in: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Berlin, Germany, 2016, pp. 1064–1074. URL: https://aclanthology.org/P16-1101. doi:10.18653/v1/P16-1101.

[19] O. Güngör, S. Uskudarli, T. Güngör, Improving named entity recognition by jointly learning to disambiguate morphological tags, in: Proceedings of the 27th International Conference on Computational Linguistics, Association for Computational Linguistics, Santa Fe, New Mexico, USA, 2018, pp. 2082–2092. URL: https://aclanthology.org/C18-1177. doi:99.9999/woot07-S422.

[20] O. Güngör, T. Güngör, S. Üsküdarli, The effect of morphology in named entity recognition with sequence tagging, Natural Language Engineering 25 (2019) 147–169. doi:10.1017/S1351324918000281.

[21] A. Katiyar, C. Cardie, Nested named entity recognition revisited, in: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), Association for Computational Linguistics, New Orleans, Louisiana, 2018, pp. 861–871. URL: https://aclanthology.org/N18-1079. doi:10.18653/v1/N18-1079.

[22] M. G. Sohrab, M. Miwa, Deep exhaustive model for nested named entity recognition, in: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Brussels, Belgium, 2018, pp. 2843–2849. URL: https://aclanthology.org/D18-1309. doi:10.18653/v1/D18-1309.

[23] A. Goyal, V. Gupta, M. Kumar, Recent named entity recognition and classification techniques: A systematic review, Computer Science Review 29 (2018) 21–43. doi:https://doi.org/10.1016/j.cosrev.2018.06.001.

[24] T. Ruokolainen, P. Kauppinen, M. Silfverberg, K. Lindén, A Finnish news corpus for named entity recognition, Language Resources and Evaluation 54 (2019) 247–272. doi:10.1007/s10579-019-09471-7.

[25] S. Almgren, S. Pavlov, O. Mogren, Named entity recognition in Swedish health records with character-based deep bidirectional LSTMs, in: Proceedings of the Fifth Workshop on Building and Evaluating Resources for Biomedical Text Mining (BioTxtM2016), The COLING 2016 Organizing Committee, Osaka, Japan, 2016, pp. 30–39. URL: https://aclanthology.org/W16-5104.

[26] B. Johansen, Named-entity recognition for Norwegian, in: Proceedings of the 22nd Nordic Conference on Computational Linguistics, Linköping University Electronic Press, Santa Fe, New Mexico, USA, 2019, pp. 222–231. URL: https://aclanthology.org/W19-6123.

[27] L. Derczynski, C. V. Field, K. S. Bøgh, DKIE: Open source information extraction for Danish, in: S. Wintner, M. Tadić, B. Babych (Eds.), Proceedings of the Demonstrations

at the 14th Conference of the European Chapter of the Association for Computational Linguistics, Association for Computational Linguistics, Gothenburg, Sweden, 2014, pp. 61–64. doi:10.3115/v1/E14-2016.

[28] J. Luoma, M. Oinonen, M. Pyykönen, V. Laippala, S. Pyysalo, A broad-coverage corpus for finnish named entity recognition, in: Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), European Language Resources Association (ELRA), Marseille, France, 2020, pp. 4615–4624. URL: https://aclanthology.org/2020.lrec-1.567.

[29] J. Luoma, L.-H. Chang, F. Ginter, S. Pyysalo, Fine-grained named entity annotation for Finnish, in: Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa), Linköping University Electronic Press, Sweden, Reykjavik, Iceland (Online), 2021, pp. 135–144. URL: https://aclanthology.org/2021.nodalida-main.14.

[30] R. Weischedel, M. Palmer, M. Marcus, E. Hovy, S. Pradhan, L. Ramshaw, N. Xue, A. Taylor, J. Kaufman, M. Franchini, M. El-Bachouti, A. H. Robert Belvin, Ontonotes release 5.0, 2013. doi:doi.org/10.35111/xmhb-2b84.

[31] K. Kettunen, L. Löfberg, Tagging named entities in 19th century and Modern Finnish newspaper material with a Finnish semantic tagger, in: Proceedings of the 21st Nordic Conference on Computational Linguistics, Association for Computational Linguistics, Gothenburg, Sweden, 2017, pp. 29–36. URL: https://aclanthology.org/W17-0204.

[32] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, in: CoRR, volume abs/1810.04805, 2018. URL: http://arxiv.org/abs/1810.04805.

[33] A. Virtanen, J. Kanerva, R. Ilo, J. Luoma, J. Luotolahti, T. Salakoski, F. Ginter, S. Pyysalo, Multilingual is not enough: BERT for finnish, CoRR abs/1912.07076 (2019). doi:10.48550/arXiv.1912.07076. arXiv:1912.07076.

[34] W. H. McNeill, Mythistory, or Truth, Myth, History, and Historians, The American Historical Review 91 (1968) 1–10. doi:10.2307/1867232.

[35] S. Myllyniemi, Suomen historian asiakirjalähteet, 2nd. ed., Kansallisarkisto; WSOY, Helsinki, 1994.

[36] S. Decker, The silence of the archives: business history, post-colonialism and archival ethnography, Management & Organizational History 8 (2013) 155–173. doi:10.1080/17449359.2012.761491.

[37] D. Colla, A. Goy, M. Leontino, D. Magro, C. Picardi, Bringing semantics into historical archives with computer-aided rich metadata generation, J. Comput. Cult. Herit. 15 (2022). doi:10.1145/3484398.

[38] R. Grishman, B. Sundheim, Message Understanding Conference- 6: A brief history, in: COLING 1996 Volume 1: The 16th International Conference on Computational Linguistics, 1996. URL: https://aclanthology.org/C96-1079.

[39] M. Kannisto, P. Kauppinen, Digital Histories: Emergent Approaches within the New Digital History, Helsinki University Press, 2020, pp. 165–180. doi:10.2307/j.ctv1c9hpt8.15.

[40] J. Jarlbrink, All the work that makes it work: Digital methods and manual labour, in: M. O. Fridlund, Mats, P. Paju (Eds.), Digital Histories: Emergent Approaches within the New Digital History, Helsinki University Press, 2020, pp. 113–126. doi:10.2307/j.ctv1c9hpt8.12.

[41] J. Clifford, B. Alex, C. M. Coates, E. Klein, A. Watson, Geoparsing history: Locating

commodities in ten million pages of nineteenth-century sources, Historical Methods: A Journal of Quantitative and Interdisciplinary History 49 (2016) 115–131. doi:10.1080/01615440.2015.1116419.

[42] J. Clifford, West Ham and the River Lea: A Social and Environmental History of London's Industrialized Marshland, 1839-1914, University of British Columbia Press, Vancouver, 2017.

[43] S. Fields, C. L. Cole, C. Oei, A. T. Chen, Using named entity recognition and network analysis to distinguish personal networks from the social milieu in nineteenth-century Ottoman–Iraqi personal diaries, Digital Scholarship in the Humanities (2022). doi:10.1093/llc/fqac047, fqac047.

[44] P. Ihalainen, B. Janssen, J. Marjanen, V. Vaara, Building and testing a comparative interface on northwest european historical parliamentary debates: Relative term frequency analysis of british representative democracy, in: CEUR Workshop Proceedings, volume 3133 of *Digital Parliamentary Data in Action (DiPaDA 2022) workshop*, Uppsala, Sweden, 2022, pp. 52–68. URL: http://ceur-ws.org/Vol-3133/paper04.pdf.

[45] E. Soper, S. Fujimoto, Y.-Y. Yu, BART for post-correction of OCR newspaper text, in: Proceedings of the Seventh Workshop on Noisy User-generated Text (W-NUT 2021), Association for Computational Linguistics, Online, 2021, pp. 284–290. URL: https://aclanthology.org/2021.wnut-1.31. doi:10.18653/v1/2021.wnut-1.31.

[46] A. Hamdi, A. Jean-Caurant, N. Sidere, M. Coustaty, A. Doucet, An analysis of the performance of named entity recognition over ocred documents, in: 2019 ACM/IEEE Joint Conference on Digital Libraries (JCDL), Champaign, IL, USA, 2019, pp. 333–334. doi:10.1109/JCDL.2019.00057.

[47] S. Orasmaa, K. Muischnek, K. Poska, A. Edela, Named entity recognition in Estonian 19th century parish court records, in: Proceedings of the Thirteenth Language Resources and Evaluation Conference, European Language Resources Association, Marseille, France, 2022, pp. 5304–5313. URL: https://aclanthology.org/2022.lrec-1.568.

The Future of Food Computing: Deepening the Scope by Broadening the Network

Maija Kāle[1], Ramesh Jain[2]

[1] University of Latvia
[2] National Institute of Advanced Industrial Science and Technology

Abstract
In recent years, there has been a growing interest in using food data in various areas of computer science. Food, as a fundamental necessity for human sustenance, undergoes complex processes from farm to fork, involving physical, biological, intellectual and emotional dimensions. The need to ensure the effective and sustainable functioning of global food systems, while addressing human and environmental challenges, necessitates the development of food computing. But the question remains: How should food computing evolve to address fundamental issues in the food systems? While global sharing of data, information and knowledge is a way forward, we propose to increase the effectiveness of this sharing by transcending narrow disciplinary boundaries and engaging a wider range of stakeholders to address practical challenges facing human society. We argue that food computing requires a multidisciplinary approach that is expansive in both scope and effectiveness. We envision a future where food computing serves as a comprehensive computational infrastructure for all aspects of food, including its production, consumption, and impact on individuals and the environment. The field of Food Computing is highly relevant to Digital Humanities (DH) researchers. By actively engaging in this area, DH researchers can make substantial contributions and shape the trajectory of future food systems, taking into account the intricate interplay of cultural, social and ethical complexities.

Keywords
Interdisciplinarity, Future, Food, Computing, Sustainability

1. Introduction

Personal food choices and consumption play an important role in public health and have a profound impact on environmental sustainability. Obesity, type 2 diabetes and cardiovascular disease are just some of the health problems associated with the diets of today's consumers. Over 60% of all deaths are caused by lifestyles where healthy food intake is one of the dominant reasons [1]. While the impact of food on personal health is an area discussed by food and health policy makers and nutritionists around the world, another new discourse has emerged in relation to food consumption, namely the impact on planetary health or the levels of biodiversity, pollution and CO2 that influence and shape climate change and the planet's ecosystems as a whole [2]. One third of global carbon dioxide emissions are attributed to food systems, with

DHNB2023 | Sustainability: Environment - Community - Data. The 7th Digital Humanities in the Nordic and Baltic Countries Conference. Oslo – Stavanger – Bergen, Norway. March 8–10, 2023.

maija.kale@lu.lv (M. Kāle); jain@ics.uci.edu (R. Jain)
https://ej.uz/nqs4 (M. Kāle); https://ngs.ics.uci.edu/ (R. Jain)
0000-0001-7116-9338 (M. Kāle); 0000-0003-2373-4966 (R. Jain)

DHNB Publications, DHNB2023 Conference Proceedings, https://journals.uio.no/dhnbpub/issue/view/875

the largest contribution coming from agriculture and land-use activities (estimated at 71% of total emissions), while the food supply chain - transport, consumption, retail and other related processes - accounts for 29% each [3]. Both deteriorating personal and planetary health are issues that global communities will have to address and provide sustainable solutions to in the long term [4]. Moreover, in an increasingly polarised world, these challenges will also need to be addressed in terms of access to food [2] and thus the ability to lead healthy lives.

Food computing, a field to which our research belongs, is a novel, interdisciplinary and forward-looking research area that aims to improve public (and increasingly planetary) health through a better understanding of food consumers and food systems. Empowered by new technology-enabled solutions, this interdisciplinary research area is expanding within academic research concerned with food consumption, public health issues and, increasingly, the environmental well-being of the planet. To understand the reasons behind various food-related issues, food computing has taken advantage of the opportunities opened up by the web revolution: social networks, mobile networks and the Internet of Things (IoT), which allow their users to easily share food images, recipes, cooking videos or record food diaries, creating large food data sets [5]. The expansion of computational studies in this area has enabled the prediction of consumers' mental states [6], the exploration of cognitive aspects of food consumption, the investigation of the impact of language on food perception and health [7], the correlation of food satisfaction with weather data [8], and other research areas are rapidly emerging.

Food is essential to human life and passes through many physical, biological, intellectual and emotional stages on its journey from farm to fork. Food is the single most important source of survival and enjoyment for every living person, regardless of where they live or their socio-economic situation. Collaboration across sectors and disciplines, including health and social science researchers, is crucial to understanding the complexity of food choices and human behaviour. The use of food computing in the digital humanities can provide researchers with a powerful toolkit to explore the many factors that shape food perception and consumption - be it the taste or colour of food, or the dining environment and weather [9, 10, 11, 8]. It is therefore imperative to envision the future of food computation as a guiding force for further progress and solutions to emerging personal and planetary health challenges. In the following sections, we present our vision for the future of food computing.

2. The Future of Food Computing

We can divide the future of food computing into several distinct areas where progress is being made. The first area is personalisation, as described in the following chapter and it includes building the food knowledge graph, developing the personal food model, capturing the context for food and developing the food recommendation engine. All these areas of development require a multidisciplinary approach and close collaboration between researchers from different disciplines - health, nutrition, social sciences and others - to create well-functioning tools that use computer science methods to work with big data and data at the personal level. This is where the contribution of DH researchers is paramount in capturing the different complexities within food systems.

The second direction for the future of food computing concerns the senses and the extent to which different senses (olfaction, taste, haptics) can be replicated or created in virtual environments. This is a dynamically developing area in relation to food computing, human-centred computing focusing on multisensory aspects or gastrophysics, and where specific progress can be observed in integrating different senses into virtual and/or augmented reality, complementing it with aspects of smell, temperature and, increasingly, taste, rather than operating only within visual and auditory formats [10]. In the context of food and human interaction, a new conceptual model and manifesto for analysing the human-food relationship has been created, focusing on the need to explore it in an open and inclusive way [12]. All the knowledge developed in the field of computer-human interaction will serve as important contextual data, taking into account the increasingly rapid merging of analogue and digital life [13]. With the rapid growth of computer-human or human-computer related data, the research scope of DH will also grow significantly.

The third area of future food computing is the collection of methods for analysing food data. Developing new approaches to using large-scale data to better understand the food consumer is highly significant, as food computing is currently locked into narrow and specific disciplines [14]. In the future, it must become more holistic. A holistic view of data is essential because research questions related to people and food exist in different scientific disciplines, but have not been answered using the large-scale data available and the new analytical tools that informatics offers for analysis. For example, a study of multisensory experiences using social media data can provide new insights into a well-researched cognitive science topic of multisensory or multimodal experiences that has not been explored using computer science techniques of natural language processing [15]. In the future, such cross-disciplinary studies should grow in scale and scope, adding new angles to food computing research questions that have tended to develop in depth rather than in scope. Interdisciplinary research teams and multi-sectoral research networks should be formed, bringing together expertise in a wide range of food-related areas, such as personal and planetary health, nutrition and food security, cognition and multisensory experience, social class and status in food choices, and many others.

The future of food computing depends on the availability and diversity of data [16], on researchers' access to data at local, regional and global levels, and on the ability of the research community to be interdisciplinary, ensuring cross-sectorality and growth in the breadth, not just the depth, of research questions. With these prerequisites in place, the future of food computing will lie in the development of personalised food applications that provide individuals with a set of tools to easily navigate the complex and multidimensional world of food.

3. Towards Personalisation

One of the ways in which food computing is developing is through personalisation. The more longitudinal individual data we have about food and its consumption, the more opportunities we have to develop personalised suggestions that take into account the context in which the individual is placed [17]. Behavioural change aspects such as adherence to healthier diets [18] with certain rewards for the right health behaviour [19] or new wearable technologies that can help monitor and suggest specific aspects to the individual [20, 21] are developing rapidly and on a large scale. The question remains - is the quest for personalisation satisfied? We see the evolution of food computing towards personalisation as a critical future infrastructure for food. The concepts and knowledge that exist in different disciplines need to be unified and shared using computational tools.

The rationale for the development of Food Computing was to develop personalised food recommendation systems that make the most of the availability of data. A food recommendation engine suggests to a user what should be consumed under given conditions in order to maximise culinary enjoyment and health benefits while minimising negative impacts on the planet. It is well known that a recommendation system takes three inputs and provides recommendations based on these inputs. These inputs are: in-depth information about the food to be recommended, the person's food model, and the context in which the recommendation is made. Based on this, the recommendation system identifies an appropriate combination of foods for the person. We would like to emphasise an important fact here: the health of the planet is ultimately determined by what individuals eat. This means that we should educate people to think about planetary health, but to eat for personal enjoyment and health. We will discuss this in more detail below.

3.1. Food Knowledge Graph

Information about food is complex. Food is an important source of enjoyment and an important determinant of individual health. In addition, a dish can be prepared in many different ways depending on the location, the availability and choice of ingredients in that location, seasonal variations and simply local tastes. A recipe used to prepare a dish lists the ingredients and the processes used to cook the dish. A dish is the result of ingredients being processed in a particular order, using processes that may use different temperatures and pressures to change the taste and nutrition of the dish. A food knowledge graph collects all this information to determine the final nutritional and taste value to recommend to a user.

Given the importance of the environmental impact of food production and distribution, it is important that this information is also included in any future Food Knowledge Graph (FKG). This information can then be used by a food recommendation engine along with nutritional and culinary information. In addition, the FKG may be able to combat some misinformation about sustainability, food and health systems by highlighting inconsistencies or other unknown relationships where they exist. While FKG will not be a panacea for the large amount of fake news in general that is present in digital media, its development should in the long run lead to more accurate and truthful information about food and health systems, thus contributing to building trust in food, health and sustainability related information.

3.2. Personal Food Model

People enjoy different foods depending on their context, including environmental and social factors. In addition, each food has a different effect on a person's biology and mood. It is well known that what a person likes is often not what their body likes. In terms of healthy vs. unhealthy food choices, the dichotomy between health and taste is pronounced because foods that are "unhealthy" are widely associated with being "tasty" [22], and public health data show that unhealthy fast food is widely consumed around the world. This is a serious problem and a major factor in the rise of chronic disease. It is well known that most chronic diseases can be prevented, delayed and controlled by diet, yet most people find it difficult to follow the advice of their dietician. By establishing a person's personal food model, which clearly identifies the type of tastes and flavours they enjoy under different climatic and social conditions, as well as how different nutrients in a food affect them biologically and emotionally, we can establish their personal food model. This food model can then be used by the recommendation engine.

The personal food model is the highest manifestation of diversity recognition in food and health system thinking, and should lead to an increased level of trust in society, which until now has only been offered non-personalised recommendations. Person-specific physiological responses to food will be predicted by compiling a variety of individualised data [23]. The next step will be to ensure that implementation can follow the personalised food model suggestions in terms of accessibility of recommended foods.

3.3. Context for Food

Many contextual factors play an important role in people's decisions about what to eat at a particular time or at a particular meal. The weather, the company or social occasion, recent meals, recent exercise and health factors, the cost of the meal and availability all play an important role. All these factors should be taken into account by a recommendation engine. However, the most important factor is the availability of food. Not everything is available at the time and place of the meal. What is available at a particular time and place is very important in deciding what could and should be consumed. Information on food availability could be compiled from several sources into a World Food Atlas. This WFA will either provide information on how to obtain foods and their characteristics in the vicinity, or where the ingredients might be available to use a particular recipe and prepare the food that could be consumed.

The other term for context for food is diversity, because instead of a single context, we live in a world of multiple contexts shaped by a myriad of social, economic and environmental factors. To capture the essence of multiple contexts, recipes that reveal ingredient pairings can be useful - showcasing elements of different national cuisines [24], but the necessary temporal dynamic can be added by analysing social media [25]. It turns out that digital social networks and digital media in general are helpful in capturing the manifestation of the 'zeitgeist', as food is richly documented and discussed in multiple formats in contemporary media [26]. The contextual knowledge contribution to the WFA is therefore also to understand food trends, including the hype foods that tend to be popular for a period of time [27], including the temporal dynamics that social media analysis offers [28]. However, this is only one aspect that the context for food space should capture. All other factors, such as the urban-rural divide, social class [29] and many others should be included, as food is the most fragile, emotional and cultural consumer product [30].

3.4. Food Recommendation Engine

As we all know, enjoying food is not just enjoying a single dish, but a pleasurable meal is a combination of several elements. These items may be selected in different quantities to balance taste, flavour and nutritional requirements, while also meeting planetary health requirements. Thus, unlike many other recommendation systems, a food recommendation engine is faced with the challenge of selecting many complementary items to meet overall culinary and nutritional needs. A food recommendation engine recommends a particular combination of dishes that is optimal for a person, rather than just one dish. In short, a food recommendation engine will involve some aspects of nudging. This will be challenging, and the diversity of researchers and experts involved in shaping the principles of nudging will be crucial to creating a reliable and trustworthy nudging recommendation engine for human and planetary health. Food is one of the most noisy topics when it comes to any recommendations that include planetary health, with meat being at the centre of the food struggle that has illuminated many societal movements, such as meat shaming, that emphasise individual responsibility rather than the architecture of food systems as such [31]. These debates are highly political and food is a proxy for value clashes. Therefore, without anthropologists and political scientists on board, any well-designed nudging food recommendation engine may fail, as culture, context and a deep understanding of food system related discourses are paramount.

Furthermore, liability, user trust and privacy are integral elements of food recommendation systems that form wider health information systems [32]. Any vision for the future of health and food recommendation systems includes ethical and privacy issues that should lead to increased trust and a move away from misinformation and disinformation. Furthermore, the quest for diversity is seen as one of the most important developments towards more trustworthy health recommendation systems - with diversity being understood as the involvement of multiple stakeholders in the health system, as well as the results produced as a ranking [33]. Discussions about diversity will inevitably lead to discussions about the results produced by food recommendation engines - transparency and open access to the algorithms used is crucial to ensure trust.

4. Conclusion

Given that our societies experience so many inefficiencies when it comes to food, while food-related big data is only growing in size and variety, it is time for the research community to plant and grow new, healthy and data-driven ideas for our food and society. In other words, it is time to diversify the scope of food computing and add interdisciplinarity and diversity to its own recipe.

We have illustrated the future of food computing through the development of food knowledge graphs, personal food models, context and recommendation engines - all areas that are based on computer science methodologies, but require a highly interdisciplinary mindset to develop nuanced research questions for personalisation development. We have also considered other developments in food computing, and consider the area of human-computer interaction to be an important future area for understanding the context of a modern human operating increasingly in a virtual environment. We also illustrate the need to consider food computing from a holistic perspective and to develop new methods to analyse increasingly complex issues related to people, food and the planet.

Finally, this study contributes to the understanding of how food computing might develop in the future. It also discusses the need to develop personalised food applications that take into account the contextual factors that society imposes on individuals. In addition, we hope that our work will encourage DH researchers to explore food computing methodologies and invite them to add the necessary layers of complexity from a DH perspective in relation to the food system and understanding of food consumers.

Acknowledgement

This research is supported by the project "Strengthening of the capacity of doctoral studies at the University of Latvia within the framework of the new doctoral model”, identification No. 8.2.2.0/20/I/006.

References

[1] G. Castiglia, A. El Majjodi, F. Calò, Y. Deldjoo, F. Narducci, A. Starke, C. Trattner, Nudging towards health in a conversational food recommender system using multi-modal interactions and nutrition labels, 2022.

[2] M. Grivins, A. Halloran, M. Kāle, Eight megatrends in Nordic-Baltic food systems, Nordisk Ministerråd, -, 2020. URL: http://urn.kb.se/resolve?urn=urn:nbn:se:norden:org:diva-7127.

[3] M. Crippa, E. Solazzo, D. Guizzardi, F. Monforti, F. Tubiello, A. Leip, Food systems are responsible for a third of global anthropogenic ghg emissions, Nature Food 2 (2021) 1–12. doi:10.1038/s43016-021-00225-9.

[4] A. Rostami, N. Nagesh, A. Rahmani, R. Jain, World food atlas for food navigation, in: Proceedings of the 7th International Workshop on Multimedia Assisted Dietary Management, MADiMa ’22, Association for Computing Machinery, New York, NY, USA, 2022, p. 39–47. URL: https://doi.org/10.1145/3552484.3555748. doi:10.1145/3552484.3555748.

[5] W. Min, S. Jiang, L. Liu, Y. Rui, R. Jain, A survey on food computing, ACM Computing Surveys 52 (2019) 1–36.

[6] K. Nakamoto, S. Amano, H. Karasawa, Y. Yamakata, K. Aizawa, Prediction of mental state from food images, in: Proceedings of the 1st International Workshop on Multimedia for Cooking, Eating, and Related APPlications, CEA++ ’22, Association for Computing Machinery, New York, NY, USA, 2022, p. 21–28. URL: https://doi.org/10.1145/3552485.3554937. doi:10.1145/3552485.3554937.

[7] M. Kāle, J. Šķilters, M. Rikters, Tracing multisensory food experiences on twitter, International Journal of Food Design 6 (2021) 181–212. URL: https://www.ingentaconnect.com/content/intellect/ijfd/2021/00000006/00000002/art00003. doi:doi:10.1386/ijfd_00030_1.

[8] M. Bujisic, V. Bogicevic, H. G. Parsa, V. Jovanovic, A. Sukhu, It’s raining complaints! how weather factors drive consumer comments and word-of-mouth, Journal of Hospitality & Tourism Research 43 (2019) 656–681. URL: https://doi.org/10.1177/1096348019835600. doi:10.1177/1096348019835600. arXiv:https://doi.org/10.1177/1096348019835600.

[9] S. Bakhshi, P. Kanuparthy, E. Gilbert, Demographics, weather and online reviews: A study of restaurant recommendations, in: Proceedings of the 23rd International Conference on World Wide Web, WWW ’14, Association for Computing Machinery, New York, NY, USA, 2014, p. 443–454. URL: https://doi.org/10.1145/2566486.2568021. doi:10.1145/2566486.2568021.

[10] C. Velasco, C. Michel, C. Spence, Gastrophysics: Current approaches and future directions, International Journal of Food Design 6 (2021) 137–152.

[11] C. Spence, Explaining diurnal patterns of food consumption, Food Quality and Preference 91 (2021) 104198. URL: https://www.sciencedirect.com/science/article/pii/S0950329321000252. doi:https://doi.org/10.1016/j.foodqual.2021.104198.

[12] M. Obrist, P. Marti, C. Velasco, Y. T. Tu, T. Narumi, N. L. H. Møller, The future of computing and food: Extended abstract, in: Proceedings of the 2018 International Conference on Advanced Visual Interfaces, AVI '18, Association for Computing Machinery, New York, NY, USA, 2018. URL: https://doi.org/10.1145/3206505.3206605. doi:10.1145/3206505.3206605.

[13] T. Andersen, D. V. Byrne, Q. J. Wang, How digital food affects our analog lives: The impact of food photography on healthy eating behavior, Frontiers in Psychology 12 (2021). URL: https://www.frontiersin.org/articles/10.3389/fpsyg.2021.634261. doi:10.3389/fpsyg.2021.634261.

[14] M. Kale, E. Agbozo, Utility of large-scale recipe data in food computing, Baltic Journal of Modern Computing 9 (2021). doi:10.22364/bjmc.2021.9.2.01.

[15] M. Kāle, J. Šķilters, M. Rikters, Tracing multisensory food experiences on twitter, International Journal of Food Design 6 (2021) 181–212. doi:10.1386/ijfd_00030_1.

[16] Y. Yamakata, S. Mougiakakou, R. Jain, Cea++2022 panel - toward building a global food network, in: Proceedings of the 1st International Workshop on Multimedia for Cooking, Eating, and Related APPlications, CEA++ '22, Association for Computing Machinery, New York, NY, USA, 2022, p. 59–60. URL: https://doi.org/10.1145/3552485.3554972. doi:10.1145/3552485.3554972.

[17] R. Jain, Lifeblood of health is data, IEEE MultiMedia 29 (2022) 128–135. doi:10.1109/MMUL.2022.3151996.

[18] A. Ghosh, Medipiatto: Using ai to assess and improve mediterranean diet adherence, in: Proceedings of the 7th International Workshop on Multimedia Assisted Dietary Management, MADiMa '22, Association for Computing Machinery, New York, NY, USA, 2022, p. 27. URL: https://doi.org/10.1145/3552484.3554368. doi:10.1145/3552484.3554368.

[19] M. Blöchlinger, J. Wu, S. Mayer, K. L. Fuchs, M. Stoll, L. Bally, Automatic classification of high vs. low individual nutrition literacy levels from loyalty card data in switzerland, in: Proceedings of the 7th International Workshop on Multimedia Assisted Dietary Management, MADiMa '22, Association for Computing Machinery, New York, NY, USA, 2022, p. 71–80. URL: https://doi.org/10.1145/3552484.3555744. doi:10.1145/3552484.3555744.

[20] O. Amft, The quest towards automated dietary monitoring intervention in free-living, in: Proceedings of the 7th International Workshop on Multimedia Assisted Dietary Management, MADiMa '22, Association for Computing Machinery, New York, NY, USA, 2022, p. 1. URL: https://doi.org/10.1145/3552484.3554984. doi:10.1145/3552484.3554984.

[21] V. Papapanagiotou, A. Liapi, A. Delopoulos, Chewing detection from commercial smart-glasses, in: Proceedings of the 7th International Workshop on Multimedia Assisted Dietary Management, MADiMa '22, Association for Computing Machinery, New York, NY, USA, 2022, p. 11–16. URL: https://doi.org/10.1145/3552484.3555746. doi:10.1145/3552484.3555746.

[22] R. Mai, S. Hoffmann, J. R. Helmert, B. M. Velichkovsky, S. Zahn, D. Jaros, P. E. Schwarz, H. Rohm, Implicit food associations as obstacles to healthy nutrition: the need for fur-

ther research, The British Journal of Diabetes & Vascular Disease 11 (2011) 182–186. URL: https://doi.org/10.1177/1474651411410725. doi:10.1177/1474651411410725. arXiv:https://doi.org/10.1177/1474651411410725.

[23] A. Leshem, E. Segal, E. Elinav, The gut microbiome and individual-specific responses to diet, mSystems 5 (2020) e00665–20. URL: https://journals.asm.org/doi/abs/10.1128/mSystems.00665-20. doi:10.1128/mSystems.00665-20. arXiv:https://journals.asm.org/doi/pdf/10.1128/mSystems.00665-20.

[24] M. Kāle, E. Agbozo, Utility of large-scale recipe data in food computing, Balt. J. Mod. Comput. 9 (2021).

[25] M. Kale, A. Siddhant, R. Al-Rfou, L. Xue, N. Constant, M. Johnson, nmT5 - is parallel data still relevant for pre-training massively multilingual language models?, in: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers), Association for Computational Linguistics, Online, 2021, pp. 683–691. URL: https://aclanthology.org/2021.acl-short.87. doi:10.18653/v1/2021.acl-short.87.

[26] M. Kāle, M. Rikters, Fragmented and valuable: Following sentiment changes in food tweets, Smell, Taste, and Temperature Interfaces CHI 2021 workshop abs/2106.04903 (2021) 100332.

[27] M. Kāle, E. Agbozo, Healthy food depiction on social media: The case of kale on twitter, in: DHN Post-Proceedings, 2020.

[28] S. C. Hutchings, Y. Dixit, M. Al-Sarayreh, D. D. Torrico, C. E. Realini, S. R. Jaeger, M. M. Reis, A critical review of social media research in sensory-consumer science, Food Research International 165 (2023) 112494. URL: https://www.sciencedirect.com/science/article/pii/S096399692300039X. doi:https://doi.org/10.1016/j.foodres.2023.112494.

[29] S. M. Finn, 81Can "Taste" Be Separated from Social Class?, in: Food Fights: How History Matters to Contemporary Food Debates, University of North Carolina Press, 2019. URL: https://doi.org/10.5149/northcarolina/9781469652894.003.0005. doi:10.5149/northcarolina/9781469652894.003.0005.

[30] R. Metcalfe, Food Routes: Growing Bananas in Iceland and Other Tales from the Logistics of Eating, MIT Press, London, England, 2019.

[31] M. Mann, The New Climate War: The Fight to Take Back Our Planet, PublicAffairs, 2021. URL: https://books.google.lv/books?id=z5flDwAAQBAJ.

[32] H. Hauptmann, A. Said, C. Trattner, Research directions in recommender systems for health and well-being: A preface to the special issue, User Modeling and User-Adapted Interaction 32 (2022) 781–786. URL: https://doi.org/10.1007/s11257-022-09349-4. doi:10.1007/s11257-022-09349-4.

[33] H. Schäfer, S. Hors-Fraile, R. P. Karumur, A. Calero Valdez, A. Said, H. Torkamaan, T. Ulmer, C. Trattner, Towards health (aware) recommender systems, in: Proceedings of the 2017 International Conference on Digital Health, DH '17, Association for Computing Machinery, New York, NY, USA, 2017, p. 157–161. URL: https://doi.org/10.1145/3079452.3079499. doi:10.1145/3079452.3079499.

*August 22, 2022

The diachrony of the new political terrorism: Neologisms as discursive framing in Swedish parliamentary data 1971–2018

Daniel Brodén, Leif-Jöran Olsson, Mats Fridlund, Magnus P Ängsal and Patrik Öhberg

University of Gothenburg, Universitetsplatsen 1, Gothenburg, 405 30, Sweden

Abstract

This paper begins to unpack the framing of terrorism in the Swedish Parliament through distant reading and by chronologically extracting neologisms in a comprehensive corpus of transcripts of parliamentary debates. Combining language technology and historical contextualization, we find support for the argument that the term 'terrorism' gained much of its modern meaning around 1970. Specifically, our study points to a legislative framing of the issue of terrorism in Swedish parliamentary debate from the early 1970s and onwards. We also find a proliferation in the production of neologisms and compounds after 9/11 2001, reflecting, among other things, the rise of a more distinct counter-terrorism discourse and more 'specialized' roles and functions related to terrorism and counter-terrorism activities. The paper concludes by emphasizing the analytical benefits of tracing parliamentary discourse through neologisms as an explorative approach to identify significant patterns for further investigation.

Keywords terrorism; parliamentary data; neologisms; mixed methods; text mining

1. Introduction

This study begins unpacking the framing of terrorism in the Swedish Parliament (the *Riksdag*) through distant reading the use of new vocabulary, i.e. neologisms, in parliamentary debates, drawing upon a combination of language technology (LT) and historical analysis. Our investigation is part of the ongoing *Terrorism in Swedish Politics (SweTerror)* project [1], a major mixed methods investigation of the national parliamentary discourse on terrorism. Building on a prior paper [2] on the neologisms of the related words 'terror' and 'terrorism' in the Swedish bicameral Parliament 1867–1970, we here turn to the unicameral Parliament from 1971–2018 asking: What compounds with terror or terrorism as an element have been added to the discourse from 1971 and onwards? Through this research question, we trace such neologisms in a comprehensive corpus of the minutes from the Parliament (*Kammarens protokoll*) with an interest in exploring significant patterns in the framing of terrorism.

1.1. The contested concept of terrorism

There is no generally accepted definition of terrorism, which can be described as an 'essentially contested concept' [3]. The term has frequently been used by various actors in conflicting ways, often

The 7th Digital Humanities in the Nordic and Baltic Countries Conference (DHNB 2023), March 8-10, 2023, Norway

EMAIL: daniel.broden@gu.se (Daniel Brodén); leif-joran.olsson@gu.se (Leif-Jöran Olsson); mats.fridlund@gu.se (Mats Fridlund); magnus.pettersson.angsal@gu.se (Magnus P. Ängsal); patrik.ohberg@gu.se (Patrik Öhberg)

ORCID: 0000-0002-5914-1516 (Daniel Brodén); 0000-0001-7107-4101 (Leif-Jöran Olsson); 0000-0002-5759-0027 (Mats Fridlund); 0000-0001-5996-5067 (Magnus P. Ängsal); 0000-0002-7433-8552 (Patrik Öhberg).

DHNB Publications, DHNB2023 Conference Proceedings, https://journals.uio.no/dhnbpub/issue/view/875

to label the actions of their opponents as illegal or illegitimate and it can partly be understood as a social and cultural construct [4]. ‘One man's terrorist is another man’s freedom fighter’ is a familiar cliché, an illustrative example being how the ANC (African National Congress) for a long time was regarded as a liberation movement by many Swedes and politicians alike [5]; at the same time it was branded as a terrorist organization by the South African and the U.S. governments.

Scholars of terrorism, to a large extent, agree that terrorism gained much of its contemporary meaning in the late 1960s and early 1970s. While the term can be dated back to the French Revolution and the Reign of Terror, the modern conception of terrorism was linked to the emergence of a particular transnational threat in the form of Palestinian commandos and other militants, including the members of the West German urban guerrilla Red Army Faction (RAF). The term brought together a multitude of violent tactics that had priorly been used in armed insurgencies – hijackings, bombings, hostage-takings, assassinations, etc. – into one single category. A major reason for terrorism being conceived then as a critical problem in dire need of a label, was that it affected the foundations of the modern (Western) social order in the form of global communications and diplomatic relations [6]. Another element in the background was policy-making initiatives primarily from the U.S. to criminalize political violence, initially aimed at curbing guerrilla warfare in Latin America [7].

To navigate this conceptual complexity, we draw upon historical contextualization and a notion of terrorism as an effect of discursive practices. Here, we understand discourse as a virtual collection of intertextually entwined utterances relating to a macro-topic and as a social practice feeding into societal structures of power and agency [8]. Specifically, we approach the discursive formation of terrorism in the Swedish Parliament in terms of ‘framing’, that is conceptual, habitually formed frames of meaning-making that are constructed through and by concrete linguistic acts [9].

1.2. Disposition

As a starting point, the paper discusses results from our prior study and comments on our LT-driven approach to the extraction of neologisms in Swedish parliamentary debates and the corpus used. We then turn to our findings and discuss results that support the argument that the concept of terrorism gained its modern meanings around the year 1970. This is followed by a contextualizing discussion of the extent to which compounds in our material point to a ‘legalistic’ framing of terrorism from the 1970s and onwards. We continue by noting a proliferation in the production of neologisms and compounds after the attacks in the U.S. on 11 September 2001, reflecting an increased focus on counter-terrorism, militant Islamism and ‘specialized’ roles and functions associated with terrorism and counter-terrorism. We conclude by stressing the analytical benefits of tracing parliamentary and historical discourse through neologism as an explorative approach to identify significant patterns for further investigation.

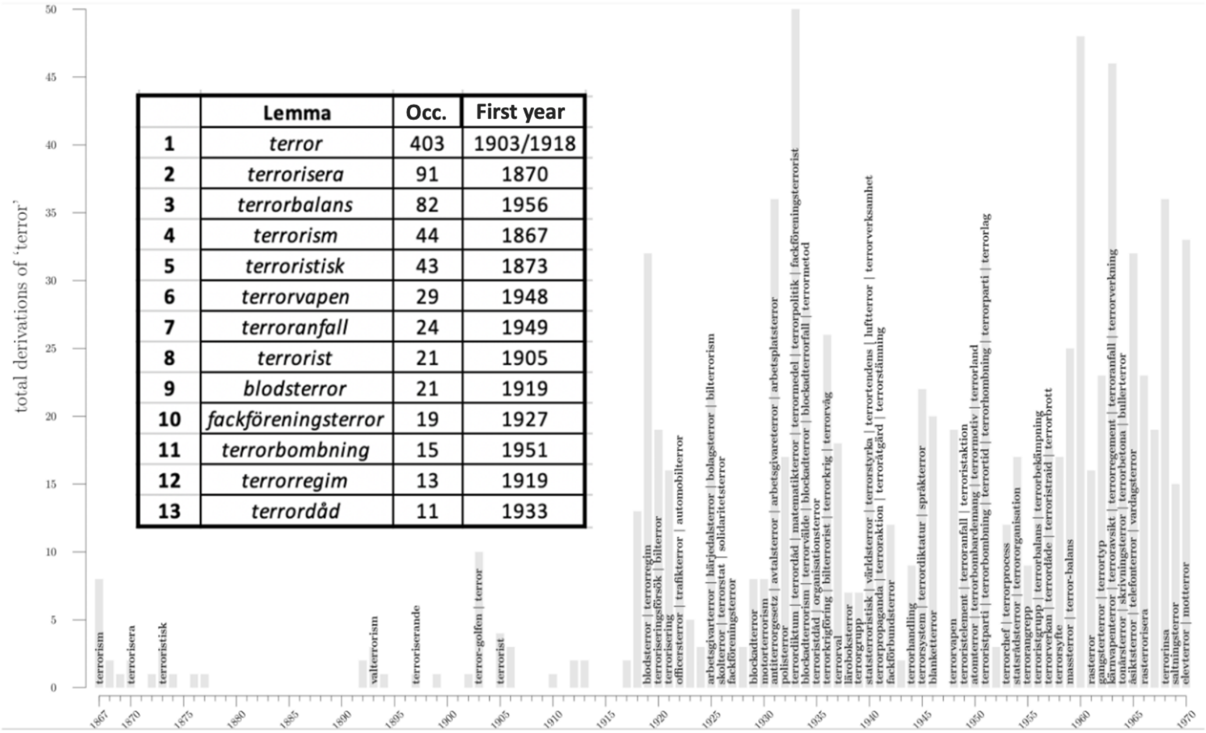
2. Parliamentary data and new vocabulary

In recent years, considerable research interest has been invested in LT-driven studies of semantic change, both concerning historical and present-day data. Internationally, much of the research has concerned word embeddings, normalization and BERT models [10] [11] [12]. At the same time, there is a growing focus on parliamentary data and the development of concepts, not least among Nordic historically-oriented scholars [13] [14] [15] [16].

2.1. Previous period

In our previous study [2], we chronologically extracted neologisms and compounds derived from the lemmas ‘terror’ and ‘terrorism’ in the transcripts of the debates in the Swedish bicameral Parliament 1867–1970 (Figure 1). This provided us with a birds-eye overview of the usage and frequency of terror and terrorism related words as well as the time of their introduction, enabling us to examine these neologisms and compounds in relation to their historical context.

Our analysis supported the argument that the terrorism gained its contemporary meanings around the year 1970 [6] [7], but also highlighted a certain complexity in its conceptual use. Among other things, we showed that *terrorism* (same spelling in Swedish and English) was used in the Swedish Parliament already in 1867, but then merely in a metaphorical way, as in the compound (*valterrorism*) in reference to perceived oppressive voting procedures in the Parliament. It was not until the early 1900s that terrorism came to denote various forms of lethal political violence, but still it was rarely used by Swedish parliamentarians (29 hits 1900–1970). Instead, other terms, including compounds consisting of *attentat* (‘attack’), and *dåd* (‘deed’), were sometimes used to label the forms of violence that we have later come to associate with terrorism. Furthermore, our study showed that the use of the word terrorism preceded that of terror (same spelling in Swedish and English). The word *terror* was first introduced as a concern for the Swedish Parliament in relation to the Finnish Civil War in 1918 and its usage primarily concerned state activities, as exemplified by compounds such as *terrorregim* (‘terror regime’) and *terrorkrig* (‘war of terror’) [17]. We could also distinguish periods of compound productivity related to domestic and geopolitical contexts. For instance, the compound *arbetsmarknadsterror* (‘labor market terror’) was used 1925–1935, denoting disruptive and at times violent actions between labor unions and employers, and *atomterror* (‘atomic terror’) 1948–1963, denoting the nuclear threat and the ‘balance of terror’ (*terrorbalansen*) during the Cold War.



	Lemma	Occ.	First year
1	*terror*	403	1903/1918
2	*terrorisera*	91	1870
3	*terrorbalans*	82	1956
4	*terrorism*	44	1867
5	*terroristisk*	43	1873
6	*terrorvapen*	29	1948
7	*terroranfall*	24	1949
8	*terrorist*	21	1905
9	*blodsterror*	21	1919
10	*fackföreningsterror*	19	1927
11	*terrorbombning*	15	1951
12	*terrorregim*	13	1919
13	*terrordåd*	11	1933

Figure 1: The production of compounds with ‘terror’ and ‘terrorism’ as an element in Swedish parliamentary debate 1867–1970. First occurrences of new derivations of terror in the bicameral corpus and staples showing the use of all derivations. The top left table shows the terror lexemes with more than 10 occurrences. Note that the frequencies are based on total occurrences, including ‘secondary’ debate text, and not ‘pure’ debate speech.

2.2. Preparatory approach – enhanced perspective

Our prior study was based on digitized Swedish parliamentary records retrieved from the National Library of Sweden. Although this data is available in fairly good OCR quality, the Westac (Welfare State Analytics) project (http://www.westac.se/en) is currently cleaning up, partly re-digitising and curating the material. The present paper is based on the latest available version (v0.4.6) of the Westac corpus of the minutes (at the time of the submission of the conference abstract) (https://github.com/welfare-state-analytics/riksdagen-corpus). There is also an ongoing exchange between Westac and SweTerror insofar as SweTerror’s LT analyst (Olsson) is further enriching and curating the data for specific research (and FAIR) purposes. Notably, we have identified minor gaps in the version of the corpus used. For instance, some 50 debate protocols have not been included for the year 1976, which only affects analyses of the corpus to a limited extent (likely there are also gaps in the digitized material from the bicameral Parliament). In total, the Westac dataset used contains 6 584 documents from the observational time (the parliamentary years 1971–2018) of which 2 062 contain the lemma terror (see below for comments about frequencies). A document contains protocols from the debates, but the number of protocols per document differs somewhat over time (while, for instance, the documents from 1979 contain between 1 and 8 protocols each, the documents from 1990 and onwards contain only 1 each). However, regardless, our analysis focuses on debates and parliamentary years rather than on documents or formalia.

In this paper, we have maintained the rudimentary yet fruitful perspective from our prior study and used a preparatory base-lining of the Westac corpus. The data was processed for analysis with tokenization, lemmatization and dependency parsing by means of the Sparv Pipeline tool designed for automatic neural and statistical annotation of documents with textual structure and linguistic properties for Swedish applications [18].

For the analysis, the debates were grouped by parliamentary year (autumn to summer, for instance 1971/1972) and queried for the lemmas ‘terror’ and ‘terrorism’. This produced data on the ‘diachrony’ of compounds with terrorism as an element, indicating the yearly production of neologisms, which was taken as the basis for our analysis. Notably, compounds that were overtly related to warfare (e.g. ‘terror bombing’), and thus outside the scope of our study, were rather unusual in the data, which also meant that they fell out of the analysis due to our criteria (see below). The distant reading was also, to a limited extent, combined with close reading to examine the specific contexts of the neologisms and compounds deemed significant, primarily focusing on the first occurrences of the words.

2.3. Pragmatic understanding of neologisms

There are different linguistic approaches to the linguistic notion of neologism. Inspired by a typology by Alexandre Rodríguez Guerra [19], we utilize a pragmatic definition that departs from the existence of a, formally, new linguistic unit or a lemma that could hitherto not be detected in a specific context, in our case the Westac corpus of the minutes. In many deliberations on the phenomenon of neologism, additional semantic properties pertaining to one and the same linguistic form, i.e. new meanings relating to one specific lexical unit, also are considered neologisms. Neologisms of this kind, however, are not possible to detect by means of the methodical approach in this study. This means, when speaking of neologisms in what follows, we thereby refer to lexical units that (1) contain either *terror* or *terrorism* as a constituent and (2) appear *in the data* for the first time. In this sense, we are, thus, dealing with neologisms in an ‘isolated’ discourse related sense, that is with relation to the specific data analyzed, albeit not with a general regard to language use in other parliamentary genres (bills, governmental reports, etc.) or, for that matter, the Swedish language community at large.

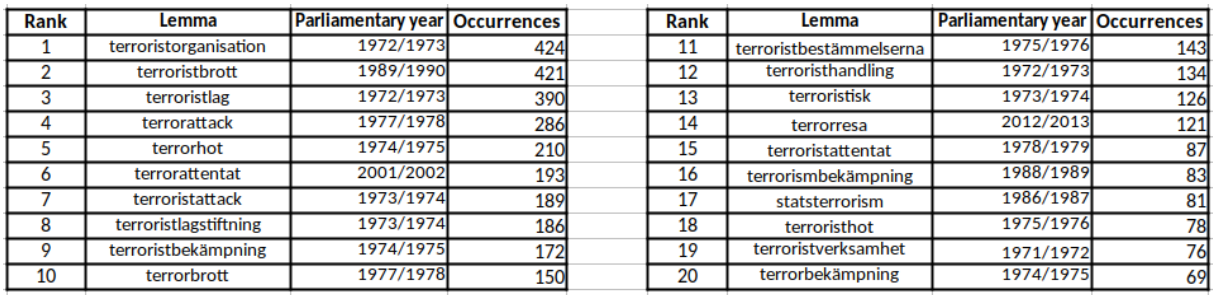
The parliamentary debates offer a rich material for understanding the use of words with terror and terrorism as an element in political discourse, since Swedish parliamentarians are free to take the floor (most ask for permission beforehand) and use them to present their own and the parties’ position on current issues [20]. While the debates are centered around the opposition’s reactions towards the government’s agenda and bills [21], they are regarded as vital for the parliamentary democratic system among all parties [22]. Notably, however, the debates are not necessarily transcribed exactly as spoken but edited by the stenographers for clarity and formality with the aim to represent the speaker’s intended meaning rather than their exact wording [23].

Since this paper concerns discourse and framing, in the discussion we will focus on words that have been used repeatedly (more than 10 times and in separate debate protocols), with a few exceptions concerning analytically interesting examples in the 2000s, keeping in mind that the time of the introduction of neologisms affects their frequencies. One should also remember that parliamentary debate speech is a particular genre and some compounds only found once in our material were likely made up on the spur of the moment [24]. Moreover, it should also be noted that the corpus not only contains debate speech but also ‘secondary’ debate text in the form of headings, results from voting, comments, etc. For instance, the compound *terroristorganisation* (‘terrorist organization) can be found

523 times in the corpus, but only 424 in actual speeches. While the use of neologisms in headings, for instance, is analytically interesting, in the following we will specifically focus neologisms found in speeches.

3. Analysis – legislation and proliferation

Terrorism became a critical political issue in Sweden in the early 1970s when militant Croatian exiles with ties to the Croatian National Resistance (HNO*, Hrvatski narodni otpor*) and the historical Ustaše movement carried out a series of attacks in the country, including the killing of the Yugoslavian ambassador in 1971 and the hijacking at Bulltofta airport in 1972. One of our more striking findings is that many of the words that are today commonly used in relation to terrorism were first introduced in the parliamentary debate in Sweden at this point (Table 1). For instance, the word *terroristverksamhet* ('terrorist activity') was first used in 1972 (76 occurrences during the parliamentary years 1971/72–2018/19), *terroristorganisation* ('terrorist organization') (424) in 1973 and *terroristattack* ('terrorist attack') (189) in 1974. This feeds into the argument that terrorism to a significant extent gained its modern meanings and usages at this point. It is worth noting that the similar compounds *terrorverksamhet* ('terror activity') and *terrororganisation* ('terror organization') had been used earlier in the bicameral debate (1940, 1954 and 1966–1968, 1970), but then almost exclusively in war-like contexts or in reference to state agents.



Rank	Lemma	Parliamentary year	Occurrences
1	terroristorganisation	1972/1973	424
2	terroristbrott	1989/1990	421
3	terroristlag	1972/1973	390
4	terrorattack	1977/1978	286
5	terrorhot	1974/1975	210
6	terrorattentat	2001/2002	193
7	terroristattack	1973/1974	189
8	terroristlagstiftning	1973/1974	186
9	terroristbekämpning	1974/1975	172
10	terrorbrott	1977/1978	150
11	terroristbestämmelserna	1975/1976	143
12	terroristhandling	1972/1973	134
13	terroristisk	1973/1974	126
14	terrorresa	2012/2013	121
15	terroristattentat	1978/1979	87
16	terrorismbekämpning	1988/1989	83
17	statsterrorism	1986/1987	81
18	terroristhot	1975/1976	78
19	terroristverksamhet	1971/1972	76
20	terrorbekämpning	1974/1975	69

Table 1: The 'top 20' neologisms with 'terror' or 'terrorism' as an element in Swedish Parliamentary debate 1971–2018.

Notably, the generic compound *terrorattentat* ('terror attack') (193) was only introduced into the parliamentary debate as late as 2001, which seems rather conspicuous in our context. However, one should note that this word was also more or less absent from the newspaper discourse up until 2001; a query in The National Library of Sweden's newspaper database (https://tidningar.kb.se) shows that the word was introduced in the Swedish press around 1990 but first came into wider use after 9/11. Other surprisingly 'late' neologisms are *statsterror* ('state terror') (11) and *statsterrorism* (state terrorism') (81) which appear in our material first in 1984 and 1986, respectively (the similar *statsterroristisk* was used only once, in 1940). [17] [25] This may seem exceptional, considering that the word terrorism has ever since its conception been associated with state repression and violence. As mentioned above, the word first appeared during the French Revolution in reference to 'the Reign of Terror'. However, the very concept of 'state terrorism' may be a more recent historical novelty. Although notions about states being perpetrators of terrorism were quite common in the 'political' 1960s and 1970s, this specific term likely found its way into the critical discourse a bit later [26].

3.1. Legalistic framing

Another major aspect of the Swedish parliamentary debate about terrorism is the extent to which neologisms and compounds have been connected to legislative concerns (Figure 2). Among the words with terror and terrorism as an element used for the first time by MPs in the early 1970s were *terroristlag* ('terrorist law') (390) in 1973; *terroristlagstiftning* ('terrorist legislation') (186) in 1974, and *terroristbestämmelser* ('terrorist regulations') (143) in 1975. It is hardly surprising that the Parliament discussed terrorism from a legislative perspective, since one of its core tasks is passing legislation. However, what is striking is that ‘legislative’ terror compounds were very rare before the 1970s and when they did appear they were used in reference to foreign events (one example being *terrorlag* (‘terror law’) apropos the situation in Europe and South Africa in 1951 and 1961, respectively). Essentially, what the sudden productivity in legalistic neologisms in the first part of the 1970s shows is that terrorism had become understood as a national problem to be dealt with by means of legislation. The killing of the Yugoslavian ambassador and the Bulltofta hijacking contributed to the Swedish Parliament adopting the Terrorist Act in 1973, aimed at preventing political acts of violence of an international nature and exclusively directed against foreign citizens suspected of ties to militant groups [27] [28].

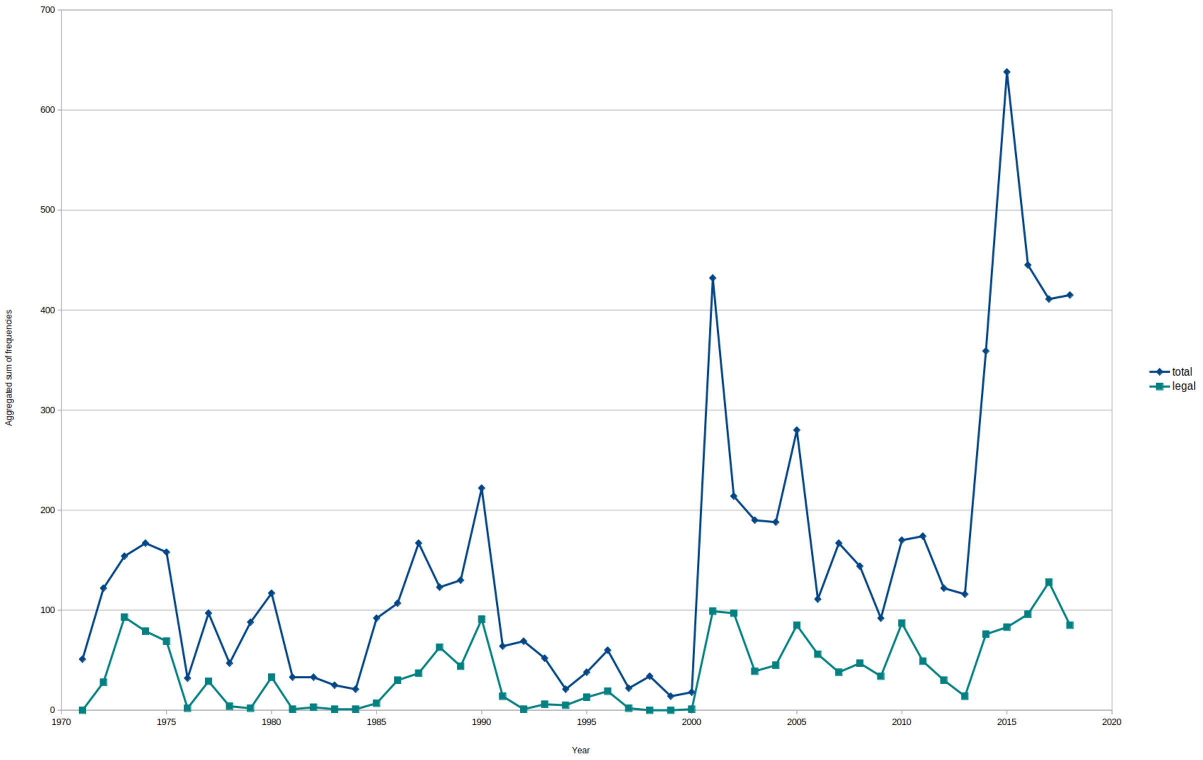


Figure 2: The frequency of compounds with ‘terror’ and ‘terrorism’ as an element in Swedish parliamentary debate 1971–2018, focusing on the relation between the total number of compounds (blue line) and ‘legalistic’ compounds (green).

This development followed an international trend insofar as terrorism was to a considerable extent understood and handled as a legal problem by Western states in the 1970s, a key issue being how international and national law might regulate acts of political violence [6] [7]. In Sweden, the parliamentary discourse about terrorism became, to a large extent, centered around the Terrorist Act for decades. This controversial legislation remained an emergency powers act during its existence that had to be renewed every year, and thus became the subject of an annually recurring parliamentary debate [27] [28]. For instance, there was a increase in the use of ‘terrorist law’ in the later part of the 1980s, reflecting the controversy surrounding the protracted municipal arrest of a number of Kurds suspected for the murders of two defectors from the PKK (Kurdistan’s Workers Party) in 1984 and 1985 as well as the assassination of prime minister Olof Palme in 1986. Overall, there was a rather consistent production of legalistic neologisms, including *terroristlista* (‘terrorist list’) (29), introduced in 1975, that referred to an official list of militant foreign organizations deemed a particular threat to Sweden. Other similar ones were *terrorbrott* (‘terror crime’) (150) in 1977, *terroristmisstänkt* (‘terrorist suspect’) (16) and *terroristbrott* (‘terrorist crime’) (421) in 1989. In fact, there was a significant increase in the use of such compounds following 9/11, partially related to the Parliament’s adoption of a number of EU counter-terrorism legislation initiatives.

3.2. After 9/11

It comes as no surprise that our results show the impact of the 9/11 attacks in the U.S. (2001) upon the parliamentary discourse on terrorism. The attacks on 9/11 ushered in a new era of concerns and in the 2000s Sweden witnessed a number of incidents related to Islamist violence, including the attempt on the life of artist Lars Vilks in the lethal Copenhagen shootings in 2015 and the lethal truck attack on Drottninggatan in Stockholm in 2017. An overview of the most frequently used words with terror and terrorism as an element *overall*, including *'old'* neologisms introduced prior to 1971, not least 'terrorism' 'terrorist' and 'terror', clearly shows the traction that the issue of terrorism gained in the parliamentary debate after 2001 (Figure 3). Whereas 97 neologisms were produced during the parliamentary years 1971/1972–2000/2001, just as many appeared during the much shorter period of 2001/2002–2007/2008. In this context, it should be noted that the lemma terror is in general more frequently occurring in the documents from the 2000s than before.

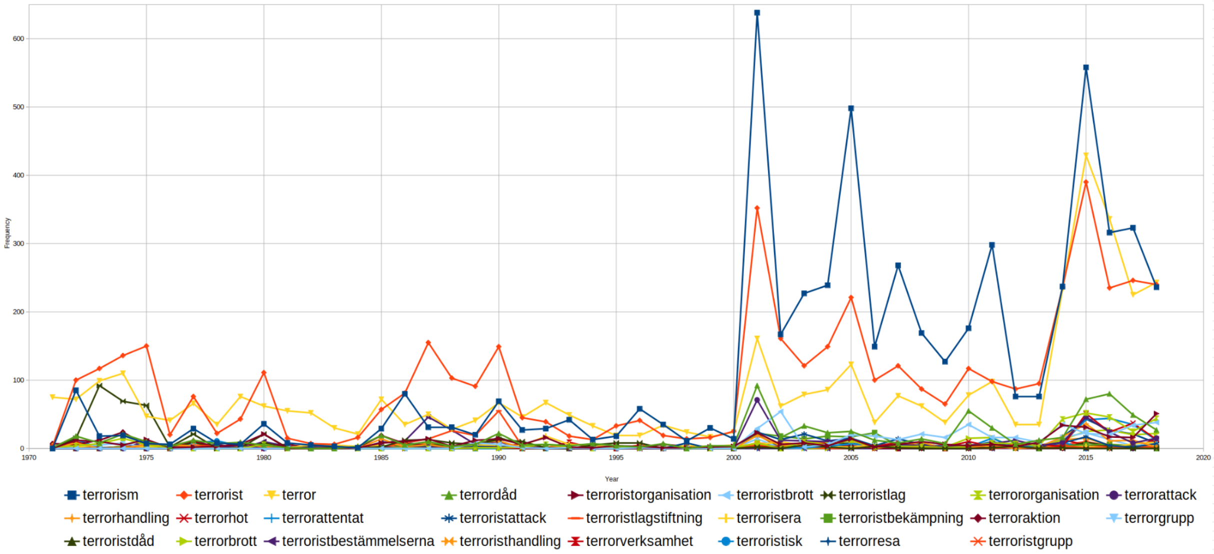


Figure 3: The frequency of compounds with 'terror' and 'terrorism' as an element in Swedish parliamentary debate 1971–2018 (lemmas with more than 100 occurrences in the corpus), including neologisms introduced prior to the period, such as the major trend lines of the 2000s 'terrorism', 'terrorist' and 'terror'.

Some of the neologisms in the 2000s were variations of familiar compounds, such as *terrorlagstiftning* ('terror legislation') (33) introduced in 2001 and 'terrorism legislation' (*terrorismlagstiftning*) (32) in 2008, but others had a more distinctly novel character. Among other things, the impact of 9/11 and the new structures of terrorism in a globalized network society [29] can be sensed in *terroristnätverk* ('terrorist network') (42) and *terrornätverk* ('terror network') (29) that both came into use in 2001. These two terms also were to a significant extent used in reference to the al-Qaida network, responsible for the attacks in the United States.

From 2014 and onwards, terrorism became associated with a particular group in a way that it had not been before. Prior, various organizations had, of course, been mentioned in parliamentary debates about terrorism, but not in the form of compounds, the few exceptions being *ustasjaterrorister* ('Ustaše terrorists') and *arabterrorist* ('arab terrorist') used once in 1971 and 1973, respectively. However, beginning in 2014, *IS-terrorist* ('IS terrorist') was used as many as 60 times in the Swedish parliamentary debate in reference to the Islamic State (IS), that is the militant salafist organization that proclaimed itself to be a worldwide caliphate, conquering areas in Iraq and Syria. The same year, *terrorsekt* ('terror sect') (34) was introduced in reference to IS. Moreover, the phenomenon of foreign fighters was also strongly associated with the organization's activities. A number of neologisms in the mid-2010s concerned persons who travel abroad to take part in armed conflicts. Already in 2013 *terrorresa* (’terror travel’) (121) appeared and was followed by the similar compounds *terroristresa* (’terrorist travel’) (26) in 2014 and *terrorismresa* (‘terrorism travel’ (45) and *terroriststridande* (‘terrorist combatant’) (19) in 2015.

Moreover, the compound terrorist combatant seems indicative of a tendency towards an explicit association of terrorism with warfare. Such a connection between terrorism and warfare could, of course, be found pre-1970, but also in the Western discourse from the late 1970s and early 1980s and onwards, for instance, in the rhetoric by the Reagan administration about terrorism as declarations of war as well as the metaphor ‘war on terror’ or the ‘global war on terrorism’ (GWOT) coined by the Bush administration for its global counterterrorism military campaign. However, compared with the US and many other European countries, in Sweden, the military discourse in policy-making on terrorism was for a long time weak. Nevertheless, at least to some extent, a conceptualisation of terrorism as a military threat started to manifest itself in the parliamentary debate in the mid 2000s through compounds such as *terroristkrig* (‘terrorist war’) (7) in 2005 and *terrorkrigsbrott* (‘terror war crime’) (6) in 2015.

3.3. Specialized roles and functions

In the 2000s, we can also see a growth in neologisms for roles, functions and stakeholders related to terrorism. Above, we noted that ‘terrorist network’ and ‘terror network’ entered the debate in 2001, and we can see other more specialized functions, such as *terrorfinansiering* (‘terror funding’) (12) in 2004 and *terrorismfinansiering* (‘terrorism funding’) (9) in 2008. One can also compare with the phenomenon of so-called foreign fighters.

At the same time, an increase in neologisms related to counter-terrorism point to the formation of a stronger counter-terrorism discourse (Figure 4). While both *terrorbekämpning* (‘terror combatting’, approx. ‘anti-terror measures’) (172) and *terroristbekämpning* (‘terrorist combatting’, approx. anti-terrorist measures’) (69) was already used in 1975 in the parliamentary debate; there has been a rise in the productivity of compounds concerning counter-terrorism since the late 1980s. For a long time, Sweden, contrary to many other Western countries, was reluctant to form a national tactical anti-terrorist strike force. Such initiatives first gained traction after the Palme killing in 1986 and the neologism *terroriststyrka* (‘terrorist force’, approx. ‘anti-terrorist force’) (6) appeared in 1989, the year that the Parliament green-lighted the creation of a national police tactical and anti-terrorist unit. In the 2000s, we also see a more Anglo-American-influenced policy-making terminology, as exemplified by the introduction of the specific term kontra*terrorism* (‘*counter*terrorism’) (13) in 2006 (c.f., priorly used terms, such as ‘*anti* terrorism’) [30]. The rise of a more distinct counter-terrorism discourse can also be discerned in neologisms referring to professionalized forms of expertise, including *terrorexpert* (’terror expert’) (9) in 2008.

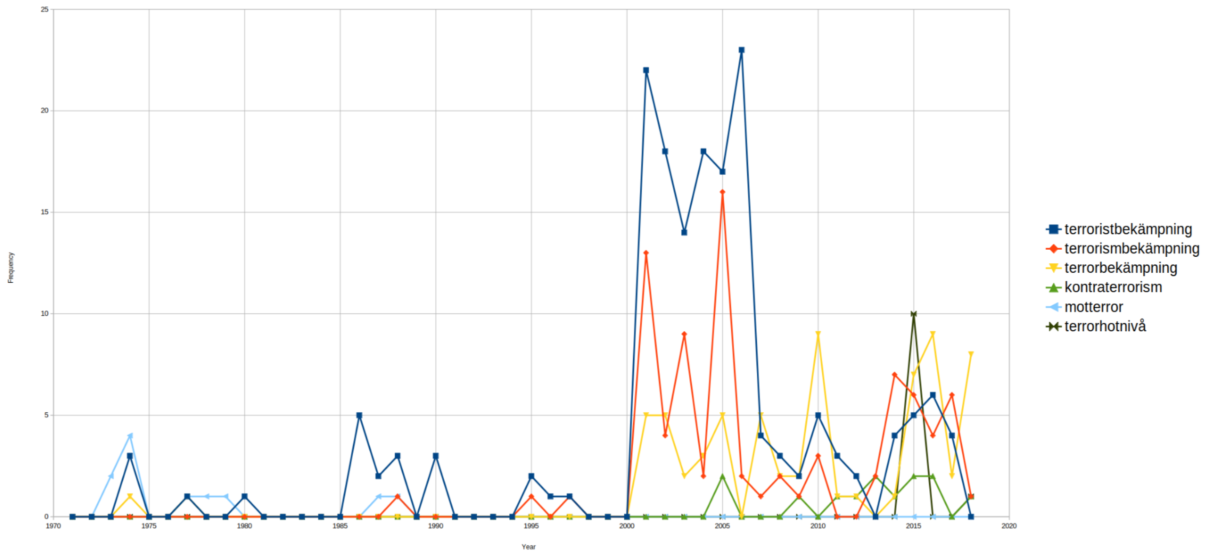


Figure 4: The frequency of the compounds related to counter-terrorism practices in Swedish parliamentary debate 1971–2018 (lemmas with more than 10 occurrences in the corpus).

Neologisms in the 2010s also seem to reflect an institutionalization of a ‘terrorismmindedness’, that is an integrated perception and practice treating terrorism as an ever-present threat [31]. The fact that the compound ‘terror threat’ (*terrorhot*) (210) was introduced already in 1974 shows that an awareness of terrorism as a potential threat was hardly new. However, in the 2010s we find neologisms that indicate a more integrated terrorismmindedness, including ‘terror threat assessment’ *(terrorhotbedömning*) (4) in 2011 and ’terror threat level’ *(terrorhotnivå*) (11) in 2016. The former compound appeared as a part of the title of the National Centre for Terrorist Threat Assessment (*Nationellt centrum för terrorhotbedömning*) established at the security service. The contrived ‘counter-terror response capability’ (*terrorbekämpningsförmåga*) (9) appearing in 2016 is another term that indicates an interest among Swedish MPs in institutionalized counterterrorism mechanisms.

4. Conclusions

This paper has studied the framing of terrorism in the parliamentary debate in Sweden by chronologically tracing neologisms and compounds in our material through distant reading. Our results, among other things, support the argument that terrorism gained its more modern meanings in the early 1970s. We also found a distinct legislative framing of the issue of terrorism that has continued over the years. The paper has also highlighted 2001 as a watershed year as there was a proliferation in the production of neologisms after 9/11, reflecting the rise of the ‘terror networks’ al-Qaida and, in particular, IS, as well as a distinct counterterrorism discourse. Furthermore, neologisms and compounds indicate a specialization of roles and functions associated with terrorism, both concerning terrorism and counterterrorism.

A main point of this paper is that even a rudimentary explorative LT approach to the extraction of neologisms in parliamentary data may yield significant results. Our findings both support previous research on the discourse on terrorism and provide historical perspectives that call for further investigation. Although this approach has already proven fruitful, a more sophisticated analysis will most likely generate more robust and detailed results. For instance, the application of word vectors will allow us to more deeply examine conceptual developments and other terms that carry similar meaning as the neologisms in our study. Also taking into account meta-data about party affiliation of speakers who use neologisms, will provide data about the extent to which different political parties have used different compounds, which allows for a more multi-dimensional analysis of the politics of terrorism. Moreover, by also focusing on close reading we can examine the context of the compounds in a more nuanced way and deepen the historical contextualization. In this sense, our approach in this paper can be understood as a vital, although rudimentary, first step towards a more complex exploration of the political and historical discourse on terrorism in Sweden.

5. Acknowledgements

The SweTerror project is funded by the research program Digitisation and Accessibility to Cultural Heritage Collections (DIGARV), which promotes data-driven research in the Humanities and Social Sciences, is jointly organized by the Swedish Research Council, the Swedish Foundation for Humanities and Social Sciences and the Royal Swedish Academy of Letters, History and Antiquities (https://www.digarv.se/en/). This work is also supported by the national e-infrastructure The National Language Bank and Swe-Clarin funded partly by the Swedish Research Council (under contract no. 2017-00626). SweTerror also draws upon the ongoing work with the Swedish parliamentary data in the Westac project (http://www.westac.se/en).

6. References

1. Edlund, Jens, Daniel Brodén, Mats Fridlund, Cecilia Lindhé, Leif-Jöran Olsson, Magnus P. Ängsal, Patrik Öhberg. 2022. A multimodal digital humanities study of terrorism in Swedish politics: an interdisciplinary mixed methods project on the configuration of terrorism in

parliamentary debates, legislation, and policy networks 1968–2018. In Kohei Arai (ed), Intelligent Systems and Applications: IntelliSys 2021. Lecture Notes in Networks and Systems (Vol. 295). Cham: Springer.

2. Fridlund, Mats, Daniel Brodén, Victor Wåhlstrand Skärström. 2022. The diachrony of political terror: Tracing terror and terrorism in Swedish parliamentary data 1867–1970. In Elena Volodina, Dana Dannélls, Aleksandrs Berdicevskis, Magnus Forsberg, Shafqat Virk (eds). Live and learn: Festschrift in honor of Lars Borin, Gothenburg: Research Reports from the Department of Swedish, Multilingualism, Language Technology.
3. Weinberg, Leonard, Ami Pedahzur, Sivan Hirsch-Hoefler. 2004. The challenges of conceptualizing terrorism. Terrorism and political violence. 16(4).
4. Brulin, Remi. 2015: Compartmentalization, contexts of speech and the Israeli origins of the American discourse on 'terrorism'. Dialectical anthropology, 39.
5. Berg, Annika, Urban Lundberg, Mattias Tydén. 2021. En svindlande uppgift: Sverige och biståndet 1945–1975. Ordfront.
6. Stampnitzky, Lisa. 2013. Disciplining terror: How experts invented 'terrorism'. Cambridge University Press.
7. Zoller, Silke. 2021. To deter and punish: Global collaboration against terrorism in the 1970s. Columbia University Press.
8. Reisigl, Martin, Ruth Wodak. 2009. The discourse-historical approach. In Wodak, Ruth, Michael Meyer (eds). Methods of critical discourse analysis. Sage.
9. Entman, Robert M. 1993. Framing. Journal of Communication 43(4).
10. Renouf, Antoinette. 2007. Tracing lexical productivity and creativity in the British media. In Munat, Judith. (ed) Lexical creativity, texts and contexts. John Benjamins.
11. Grieve, Jack, Andrea Nini, Diansheng Guo. 2017. Analyzing lexical emergence in Modern American English online. English language and linguistics. 21(1).
12. Säily, Tanja, Eetu Mäkelä, Mika Hämäläinen. 2021. From plenipotentiary to puddingless: Users and uses of new words in Early English Letters. In Mika Hämäläinen, Niko Partanen, Khalid Alnajjar (eds). Multilingual Facilitation. Rootroo.
13. Marjanen, Jani, Jussi Kurunmäki. 2020. How ideology became isms. In Heikki Haara, Koen Stapelbroek & Mikko Immanen (eds). Passions, politics and the limits of society. Helsinki Yearbook of Intellectual History, 1. de Gruyter.
14. Jarlbrink, Joham, Fredrik Norén, Robin Saberi. 2022. Contextual modeling of 'propaganda', 'information' and 'upplysning' in Swedish parliamentary speeches, 1920–2019. Digital parliamentary data in action (DiPaDA 2022) workshop, Uppsala University, Sweden, March 15, 2022.
15. Norén, Fredrik, Johan Jarlbrink, Alexandra Borg, Erik Edoff, Måns Magnusson. 2022. The transformation of 'the political' in post-war Sweden. In Estelle Bunout, Maud Ehrman, Frédéric Clavert (eds). A new Eldorado for historians? Reflections on tools, methods and epistemology. De Gruyter.
16. Ohlsson, Claes, Victor Wåhlstrand Skärström, Henrik Björck. 2022: The market as a concept in Swedish parliamentary records from 1867 to 1970: A mixed methods study. Digital Parliamentary Data in Action (DiPaDA 2022) workshop, Uppsala University, Sweden, March 15, 2022.
17. Fridlund, Mats, Daniel Brodén, Leif-Jöran Olsson, Magnus P Ängsal. 2022. Codifying the debates of the Riksdag: towards a framework for semi-automatic annotation of Swedish parliamentary discourse'. In Matti La Mela, Fredrik Norén, Eero Hyvönen. (eds). DiPaDa 2022: Proceedings of the Digital Parliamentary Data in Action (DiPaDa 2022) Workshop.
18. Borin, Lars, Markus Forsberg, Martin Hammarstedt, Dan Rosén, Roland Schäfer, Anne Schumacher. 2016. Sparv: Språkbankens corpus annotation pipeline infrastructure. In The Sixth Swedish Language Technology Conference (SLTC). Umeå University.
19. Rodríguez Guerra, Alexandre. 2016. Dictionaries of neologisms: A review and proposals for its improvements- Open Linguistics, 2(1).
20. Bäck, Hanna, Marc Debus. 2016. Political parties, parliaments and legislative speechmaking. Springer.

21. Helms, Ludger. 2008. Studying parliamentary opposition in old and new democracies: Issues and perspectives. The journal of legislative studies, 14(1-2).
22. Proksch, Sven-Oliver, Jonathan B Slapin. 2015. The politics of parliamentary debate: parties, rebels and representation. Cambridge University Press.
23. Thelander, Kerstin. 1986. Politikerspråk i könsperspektiv, Malmö.; Hallberg, Anita. 1994. Kan man forska på riksdagsprotokollet?. Språkvård, 30(3).
24. Laver, Michael, Kenneth Benoit, John Garry. 2003. Extracting policy positions from political texts using words as data. American political science review. 97:2.
25. Ängsal, Magnus P, Daniel Brodén, Mats Fridlund, Leif-Jöran Olsson, Patrik Öhberg. 2022. Linguistic framing of political terror: Distant and close readings of the discourse on terrorism in the Swedish Parliament 1993–2018, CLARIN annual conference proceedings, 10–12 October 2022, Prague.
26. Jarvis, Lee, Michael Lister. 2014. State terrorism research and critical terrorism studies: An assessment. Critical Studies on Terrorism 7(1).
27. Ribbing, Antonia. 2000. Sveriges terroristbestämmelser – brottsprevention och demokratiska rättsstatsideal. In Janne Flyghed (ed). Brottsbekämpning – mellan effektivitet och integritet: Kriminologiska perspektiv på polismetoder och personlig integritet. Studentlitteratur.
28. Hansén, Dan. 2007. Crisis and perspectives on policy change: Swedish counter-terrorism policymaking. Diss. Swedish Defense University.
29. Castells, Manuel. 1997. The power of identity: information age: Economy, society and culture, Volume II, Blackwell.
30. On Swedish political agenda-setting on national security issues in general, see Eriksson, Johan. 2004. Kampen om hotbilden: Rutin och drama i svensk säkerhetspolitik. Santérus.
31. Fridlund, Mats. 2011. Buckets, bollards and bombs: Towards subject histories of technologies and terror. History and Technology. 27.

The Cultural Imaginary of Terrorism: Close and Distant Readings of Political Terror in Swedish News and Fiction During the Cold War

Mats Fridlund, Michael Azar, Daniel Brodén and Michael McGuire

University of Gothenburg, Universitetsplatsen 1, Gothenburg, 405 30, Sweden

Abstract

The digital history project 'The Cultural Imaginary of Terrorism' (2022–2025) examines the cultural meaning-making of political terror in Swedish nonfiction and fiction during the Cold War, a critical period for the formation of the international discourse on terrorism. To explore the Swedish 'cultural imaginary of terrorism' is to study how a society makes sense of terrorism and itself in relation to the phenomenon through figures of thought, frames of reference and fantasies in, among other things, national newspapers and periodicals. This paper gives an overview of the project, demonstrating our integrative use of distant and close reading methods through a pilot study of elements of the cultural imaginary of terrorism as represented in the conservative periodical *Svensk Tidskrift* 1945–1991. Our exploratory analysis focuses on the extraction of relevant texts, the visualization and mapping of the discourse through the development of key terms as well as individuals, places, groups and states associated with political terror and terrorism. We conclude by stressing the benefits of an integrative research design drawing upon complementary perspectives: how our text mining methods allow us to identify significant patterns in the text data and how our historical expertise allows us to single out aspects that call for further investigation.

Keywords terrorism, cultural imaginary, text mining, digital history, mixed-methods

1. Introduction

This scoping paper presents the core approach and elements of a digital history project that constitutes an integrative digital humanities (DH) effort geared towards synergetic and enhanced humanities knowledge production, rather than an investigation of the development, applicability or evaluation of tools and methods. *The Cultural Imaginary of Terrorism* (2022–2025) combines close and distant reading approaches in an integrative interdisciplinary study of cultural engagement with political terror in Sweden during the Cold War. Here, we present the design of this project, emphasizing how a collaborative approach to the use of standard text mining and data visualization methods can enhance our understanding of a historical process. On a higher level, the paper feeds into the debates within DH about the necessity of pursuing intersections of interpretative and computational analysis through a collaborative process where humanities scholars and data analysts are in a dialectical relationship [1, 2].

The project brings together scholars with proficiency in digital history, DH and corpus linguistics as well as with domain expertise in terrorism studies, intellectual history and media history to answer the overarching research question: *How was terrorism depicted in Swedish media and culture during the*

The 7th Digital Humanities in the Nordic and Baltic Countries Conference (DHNB 2023), March 8–10, 2023, Norway

EMAIL: michael.azar@gu.se (Michael Azar); daniel.broden@gu.se (Daniel Brodén); mats.fridlund@gu.se (Mats Fridlund); michael.mcguire@gu.se (Michael McGuire)

ORCID: 0000-0002-5914-1516 (Daniel Brodén); 0000-0002-5759-0027 (Mats Fridlund); 0000-0002-7348-0127 (Michael McGuire)

DHNB Publications, DHNB2023 Conference Proceedings, https://journals.uio.no/dhnbpub/issue/view/875

Cold War? Our study is augmented with the parallel running integrated mixed-methods project *Terrorism in Swedish politics* (SweTerror) (2021–2024) [3], drawing upon state-of-the-art multimodal Language Technology (LT) approaches and mapping the parliamentary debate, partly during the same period (1968–2018).

Although Sweden was less affected by political terror than many other countries in Western Europe, terrorism became a cause for national concern in the 1970s (Figure 1), following a number of political assassinations and hostage takings, including the killing of the Yugoslavian ambassador in 1971 and the Bulltofta hijacking in 1972, both carried out by militants associated with the Croatian National Resistance (HNO*, Hrvatski narodni otpor*) and the historical Ustaše movement. As lethal political violence had been virtually unheard of in postwar Sweden, the Parliament felt compelled to adopt a controversial counter-terrorism law in 1973, directed against foreign citizens with ties to militant organizations. Terrorism remained in the public eye in the following decades, through domestic incidents such as the occupation of the West German embassy, during which members of the RAF (*Rote Armee Fraktion*) executed two hostages before accidentally blowing up the building, and a foiled retaliatory plot to kidnap former Minister Anna-Greta Leijon, involving both foreign militants and young Swedes. In the 1980s, the protracted 'municipal arrests' of a number of Kurds, suspected for involvement in the murders of two defectors from the PKK in 1984 and 1985, as well as the killing of prime minister Olof Palme in 1986, garnered considerable attention in the media.

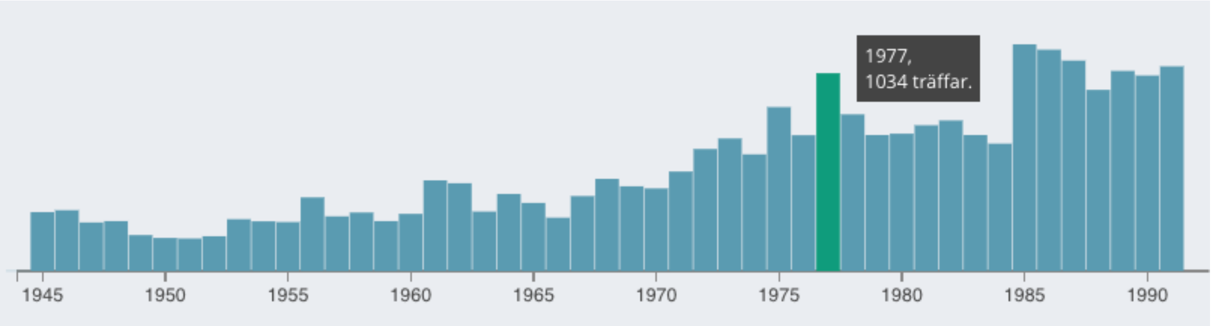


Figure 1: Search profile for '*terror*' in *Dagens Nyheter* during the Cold War (tidningar.kb.se). The 1977 peak is primarily connected to West German terrorism.

This paper begins by situating the project within the wider context of research on the history of terrorism. Next, we turn to the integrative DH approach and present our analytical framework, primary material and methodological approach. We describe our scope and focus on tracing the cultural meaning-making of political terror in national newspapers, periodicals, works of fiction and nonfiction, and our use of text mining to map the meanings attached to terrorism in the discourse. This is followed by a more detailed presentation of a pilot study in the form of an exploratory distant and close reading of the conservative periodical *Svensk Tidskrift*, serving as a prototype for our approach. We conclude by drawing attention to some preliminary results and affordances provided by our integrative DH approach.

2. The Domain: Terrorism in Cold War Sweden

The study of terrorism as an academic research domain goes back to the 1970s, when 'terrorism' became the preferred label for a range of acts of political violence, partly as a product of the Cold War [4, 5]. Initially, this field, recently often referred to as Orthodox Terrorism Studies (OTS), suffered from its closeness to state interests, but this changed with its rapid post-2001 growth, the emergence of Critical Terrorism Studies (CTS) and more self-reflexive and critical approaches [6, 7, 8]. Our understanding of terrorism follows the CTS 'minimal foundationalist definition': 'violence or its threat intended as a symbolically communicative act in which the direct victims of the action is instrumentalized as a means to creating a psychological effect of intimidation and fear in a target audience for a political objective' [9]. Essentially, CTS analyzes terrorism as discursive processes; how acts and actors of political violence are socially constructed as 'terrorism' and 'terrorists' as well as the role of state terrorism, which is excluded from many general OTS-definitions [10]. However, only rarely have terrorism scholars studied the conceptual development of terrorism through quantitative and data-rich approaches, cf. [11, 12, 13, 14, 15].

Notably, there have also been few elaborated historical studies about terrorism in Sweden. Some research has covered specific terrorism-related events before the Cold War, including the Amalthea bombing in 1908 and other violent incidents related to the early history of the labor movement [16, 17, 18], or touched on the topic [19, 20, 21, 22]. Political scientists and criminologists have, to some extent, studied Swedish counter-terrorism policymaking, focusing on legislation and intelligence communities [23, 24, 25, 26], but, there are virtually no dedicated studies of terrorism as a theme in Swedish literature, culture and debate (for partial exceptions, see [27, 28, 29, 30]). Consequently, *The Cultural Imaginary of Terrorism* will contribute to filling a gap in research on the discourse of terrorism in Sweden, empirically as well as theoretically.

Following Stampnitzky's [4] argument that the modern understanding of terrorism is partly a product of the Cold War, we study the development of the cultural imaginary of terrorism in Swedish media and culture during the period, exploring the significance of the geopolitical tensions between east and west, north and south, left and right, etc. Our scope covers the 'exceptionalism' of the *Folkhemmet* era as well as the political radicalism of the 1960s and 1970s and the period leading up to the end of the Cold War. Central to the study is how Sweden's 'Third Way' policy contributed to a self-image associated with peacefulness, neutrality and rationality in contrast to a conflict-ridden world [31]. We also use a post-colonial lens to explore how geopolitical tensions between the USA and the Soviet Union fed into the struggles for independence in former colonies and the post-1968 emergence of political violence in Western Europe that came to be known as terrorism. Thus, the project will encompass several debates and controversies in Sweden surrounding foreign militant organizations (Ustaše, RAF, PKK, etc.) and struggles (the Israeli-Palestinian conflict, the Algerian War, etc.), depicted in print media as well as works of fiction and nonfiction on terrorism.

A critical part of the study is devoted to the different, at times contradictory, perspectives on terrorism in Swedish media and culture, including the extent to which terrorism has been framed as a 'foreign' or a 'domestic' phenomenon and how different forms of terror-related violence have been explained and demarcated by different actors and in different contexts. Our analysis pays attention to the definitions that have been employed for framing terrorism and the alternative terms used ('resistance struggle', 'guerrilla warfare', 'state terror', etc.), also drawing upon the 'minimal foundationalist definition' of terrorism (see above) to discuss neglected or 'downplayed' events. We also focus on specific stereotypes (martyr vs. fanatic, victim vs. villain, innocent vs. guilty, etc.) and tropes (good vs. evil, East vs. West, civilization vs. primitivism/barbarism, masculinity vs. femininity, rationality vs. madness, etc.) that have been involved in the representation of terror-related violence.

3. Distant History: Mixing Methods

Primarily, *The Cultural Imaginary of Terrorism* contributes to the methodologically driven field of digital history. While digital history is arguably one of the more vibrant fields within the realm of DH, in Nordic countries, it is primarily Finnish digital historians that have produced major research within the area, see [32, 33, 34, 35, 36], although interest is currently picking up in Sweden [37, 38, 39, 40]. Partly, the project draws upon standard DH approaches, including text mining and distant reading techniques for large-scale analysis of historical phenomena and processes, c.f. [41, 42, 43], but also employs an integrative mixed-methods approach. As previously noted, digital history includes a variety of digital methods with varying degrees of technical sophistication. For instance, while individual historians may use already digitized sources and off-the-shelf pre-programmed digital tools (mainly various search or information retrieval interfaces) interdisciplinary teams that includes technical expertise with programming or coding skills, often use advanced equipment or methodologies (such as GIS, language models, agent-based modeling etc.). This has been described in terms of three different methodological approaches: digital history 1.0, digital history 1.5 and digital history 2.0, depending on the historical researcher's use and consideration of advanced digital affordances and resources [44].

This project belongs both to the 'semi-automatic' digital 1.5, as it includes historical researchers performing individual research on digitized corpora using standard digital search tools and simple concordancers, and digital history 2.0 in that a substantial part of the project consists of collaborative interdisciplinary research by historians and corpus linguists.

3.1. Analytical Approach

The project approaches the cultural sense- and-meaning-making of terrorism in Sweden during the Cold War through the key concept of 'the cultural imaginary of terrorism' [45]. Rather than simply being equated with the public discourse, this concept refers to a specific culture's imaginings about that which is perceived as terrorism. The cultural imaginary has a precursor in the concept of 'the social imaginary', which has a long theoretical lineage and concerns the ways in which a society perceives itself, its history and future, see [46, 57, 48, 49, 50]. Notably, the imaginary does not necessarily imply false representations of reality. Rather, to explore the cultural imaginary of terrorism is to study how a certain culture makes meaning and makes sense of terrorism as well as itself in relation to the phenomenon. Hence, to examine the cultural imaginary as a site of both discursive convergence and conflict, we consider the existence of multiple cultural imaginar*ies* of terrorism, c.f., [51, 52].

The specific research questions in the project are grouped into three strands:

Imagining terrorism investigates imaginative thinking about terrorism as a transformative force with a focus on the relationship between terrorism, the welfare state and citizens [30]. Drawing further on the concept of 'the cultural imaginary of terrorism' [45, 53, 54], we analyze the complex interplay between actual instances of political violence and factual as well as fictional cultural representations that concerns the bearing of terrorism upon Swedish society and its past and future. *Key research questions*: In what ways have different forms of (factual and fictional) terrorism and acts of political terror been represented as an existential threat, emancipatory promise or politically irrelevant to Swedish society? In what ways have imaginative thinking and 'what if' scenarios about acts of terrorism served as instruments for fear-mongering, criticism of the existing order or progressive visions?

Domesticating terrorism explores how terrorism is established as a 'domestic' and 'normal' phenomenon through news, debate, fiction, and nonfiction. Drawing on the concept of 'terrormindedness' [55], we explore the extent to which terrorism in Cold War Sweden is conceived as a phenomenon occurring in Sweden or perpetrated by its citizens as well as an increasingly domestic and decreasingly exceptional Swedish phenomenon (cf. 'banal terrorism', [56]; 'terrorism and securitization', [57]). *Key research questions*: In what ways have (factual and fictional) incidents of terrorist violence in Sweden been perceived and represented as 'unimaginable' and 'unprecedented', 'familiar' or 'recurring'? To what extent have 'terroristic' traits of Swedish acts of political violence been repressed and historical 'Swedish' acts of political violence (the 1908 Amalthea and 1940 *Norrskensflamman* bombings, etc.) been reframed as domestic terrorism?

Exteriorizing terrorism examines the discourse of terrorism as an inherently foreign phenomenon, whether occurring in Sweden or abroad [58, 59]. Drawing on the concepts of 'othering' and 'sameness', we explore how certain perpetrators of political violence are identified as 'not one of us', culturally 'alien' and threatening, while others are acknowledged as 'one of us', sharing a way of thinking, being and aspiring [60, 61, 62]. *Key research questions*: To what extent have the motivations and ideologies of foreign citizens involved in (factual and fictional) acts of political violence been represented as 'terroristic' on the same grounds as Swedish citizens? How has terrorism been framed in relation to anti-colonial liberation and separatist struggles and conflicts supported military or politically by the Soviet Union or the USA (Middle East, Nicaragua, Yugoslavia, South Africa, etc.)?

3.2. Material

To capture salient aspects of the cultural imaginary of terrorism in Sweden during the Cold War, the project studies different empirical materials that, taken together constitute principal channels for the cultural conversation on terrorism in print media during the period.

We examine different forms of journalistic meaning-making that, to a large extent, are colored by ideologies and agendas. This meaning-making also receives much of its influence and structural coherence from different generic forms [63, 64, 65].

1) Newspaper coverage of terrorism focuses on articles in editorial, debate and culture sections in seven nationwide daily and evening papers: *Dagens Nyheter, Göteborgs-Posten, Svenska Dagbladet, Sydsvenska Dagbladet, Aftonbladet, Expressen,* and *Göteborgs-Tidningen/GT*. These newspapers represent a broad ideological spectrum and are also, to some extent, from different parts of the country.

2) The discussion of terrorism in selected cultural periodicals, including *BLM, Clarté, Folket i Bild/Kulturfront, Judisk krönika, Marxistiskt Forum*, *Ord & Bild,* and *Svensk Tidskrift*. These periodicals have been selected based on their centrality or relevance to terrorism in the public dialogue and their representation of a range of political perspectives, from left-wing to right-wing.

We also study the depiction of terrorism in literature as a subject matter for Swedish writers – left, right, and middle; highbrow, middlebrow, and popular; experimental, conventional, and formulaic [66, 67]. Our analysis includes both distant and close reading of the texts and their reception in reviews.

3) The framing of terrorism in nonfiction books on the topic (approx. 20 titles), including titles such as journalist Janerik Larsson's *Politisk terror i Sverige* ('Political terror in Sweden', 1968), jurist Göran Melander's *Terroristlagen: ett onödigt ont* ('The Terrorist Act: An Unnecessary Evil', 1975), journalist Hans Hederberg's *Operation Leo* (1978) and opinion maker Bertil Häggman's *Moskva och terroristinternationalen* ('Moscow and the Terrorist International', 1984). Primarily, we focus on books by Swedish authors, but also consider translated titles.

4) The depiction of terrorism in Swedish fiction books on the theme (approx. 70 titles), spanning a variety of genres. Prominent titles include Per Anders Fogelström's historical novel *Café Utposten* ('Café The Outpost', 1970), Sture Karlsson's non-fiction novel *Det förlorade landet* ('The Lost Land', 1974), Maj Sjöwall's and Per Wahlöö's police novel *Terroristerna* ('The Terrorists', 1975), P.O. Enquist and Anders Ehnmark's published theater play *Mannen på trottoaren* ('The Man on the Sidewalk, 1979), Carl-Henning Wijkmark's realist novel *Sista dagar* ('Last Days', 1986) and Jan Guillou's political thriller *Den demokratiske terroristen* ('The Democratic Terrorist', 1987).

By not only considering fiction and nonfiction as separate realms for imaginative thinking, but also studying how they intersect, the project directs attention to both significant differences and discursive overlaps in the development of the cultural imaginary of terrorism in Sweden. Notably, our analysis pays particular attention to key actors who traverse these different media genres. For example, Jan Guillou and Janerik Larsson wrote journalistic articles in newspapers and cultural periodicals as well as authored nonfiction and fiction on the topic of terrorism.

3.3. Distant Reading

In line with digital history's integrative approach, this project fuses close reading of empirical sources in their contexts and distant reading, aided by computational methods from DH and corpus linguistics. Text mining of national newspapers and periodicals significantly enhances the scope and validity of the study. These texts are stored at the database of the National Library of Sweden (KB) and accessed in its digital lab (KBLab) or otherwise available in digital formats. This allows us to systematically identify keywords, text segments and larger articles related to the topic of terrorism and creates overviews and analytical visualizations of discursive patterns by studying the development of key terms, such as 'terrorism', 'urban guerilla', 'state terrorism', etc. We will utilize a range of methods from corpus linguistics that enables us to: 1) search out specific persons, places and organizations (named-entity recognition); 2) quantitatively map the associations of terms such as 'terrorism', dislocations in meaning and collocations; 3) discover underlying themes (topic modeling); 4) identify changes in conceptual meanings and associations over time and across different publications (word vectors) (see [68, 69]. Notably, this analysis is informed by critical DH perspectives on the use of data-intensive methodologies [69, 70, 71].

The project also draws upon social network analysis (SNA), allowing us to identify actors (authors, reporters, interviewees, etc.) and their discursive networks that have figured prominently in media and culture. Additionally, we use discourse network analysis (DNA) [72] and controversy mapping [73] informed by actor-network theory (ANT) analysis [74, 75, 76]. Notably, we incorporate historical events and book titles in the network analysis, enabling us to establish whether certain concepts, publications and incidents (e.g., the 1972 Munich Olympics massacre, the 1975 West Germany Embassy siege in Stockholm) contributed to defining the meaning-making about terrorism in newspapers and periodicals.

4. Pilot Study: Distant Reading of a Cold War Cultural Periodical

One of the central source materials for mapping the Swedish meaning-making about terrorism is 'cultural periodicals' (*kulturtidskrifter*) and in the following we perform a pilot study, using distant reading to explore a corpus created from content of *Svensk Tidskrift* (SvT) 1945–1991. SvT is claimed to be 'Sweden's oldest political journal' and is a conservative-liberal cultural periodical published from 1911 and onwards [77] with several predecessors. The first predecessor was *Svensk literatur-tidskrift* (Swedish literary journal), published from 1865–1869, which in 1870 changed its name to *Svensk Tidskrift för litteratur, politik och ekonomi* ('Swedish journal for literature, politics and economy'), published until 1877. The periodical was resumed as *Ny Svensk Tidskrift* ('New Swedish journal'), published 1880–1890 and in 1891 it was restarted as the new journal *Svensk Tidskrift* (published until 1895) and in 1911 with the old name *Svensk Tidskrift*. The new journal was published until 1932 when it had a three-year interruption after which it was again published continuously from 1936–2004. In 2006, it was, once again, restarted as an online-only journal.

SvT was and still is close to the Swedish Conservative Party (*Högerpartiet* 1938–68, *Moderata Samlingspartiet* 1969–) and is described as the party's 'forum for ideas and an 'independent organ' for its 'academic faction'. The different editorial boards have allegedly been 'united' in their 'critique of totalitarian currents – socialist as well as fascist.' SvT represented an ideology 'characterized by a strong feeling for the importance of the European cultural tradition' and with a reformist program based on liberal conservatism. [78, 79] Several of the editorial members were, like most of the chief editors, academics in humanities and social sciences and/or conservative politicians. The chief editors during the Cold War period were as follows; the political scientist and Member of Parliament (MP) for the Conservative, Party Erik Håstad (1936–48); the political economist and former Conservative Party leader and Minister of Education, Gösta Bagge (1948–51); the legal historian Gerhard Hafström, (1951–57); the legal historian, Erik Anners (1957–80); the journalist and Conservative Party MP, Margaretha af Ugglas (1980–90); the Conservative Party MP, Gunnar Hökmark (1990); and the secretary of state in a conservative led government, Odd Eiken (1991).

4.1. Creation of Journal Article Corpus

At SvT:s centenary in 2011 a digitized archive was made publicly available on the online journal website *svensktidskrift.se*. This was in the form of references to digitized versions of all issues from 1911–1932, available at http://runeberg.org/svtidskr/ and as searchable files of all articles from 1936 onwards at https://svensktidskrift.se/arkivet/. We used the latter to scrape all articles containing terror-related words into a corpus.

During the Cold War, SvT was published continuously with 8–10 issues per year for a total of some 400–600 pages annually. Overall, the periodical contained two types of articles: signed articles providing perspectives or studies of contemporary and historical issues and unsigned editorial articles as well as shorter articles collected under the heading *Dagens frågor* ('Issues of the day') that commented on current events. Articles relating to terrorism were retrieved from the SvT website and through Google advanced search. The query terms used to retrieve articles for the period of 1945–1991 were *terror*, *terrorism*, and *terrorist*.

In total, 189 articles were retrieved after eliminating doubles and separating the terror-relevant articles of *Dagens frågor*, creating a corpus of 416.707 word tokens. While SvT provides both text (txt) files and PDF image files of previously published articles, only text files were used for creation of the corpus. The text files contain small numbers of OCR errors however, these errors were disregarded as inconsequential for this initial pilot and the statistical analysis. The corpus was annotated with part-of-speech tags and lemmatization using TreeTagger [80, 81] with the Swedish language parameter file. TXM [82, 83, 84] and AntConc [85] were both used as concordancers.

4.2. Analytical Methodology

First, we collated a list of all titles inspected for terror-related terms and a list of all author names, inspected for repeating authors, prominent individuals, known authors of terrorist fiction or nonfiction (like journalist Janerik Larsson) and other individuals also occurring in the project's other materials (like publisher Bo Cavefors).

We used semi-automated Named Entity Recognition (NER) to extract the most frequently occurring individual, organization and place names in the texts. Initially, we started by using search results for nouns beginning with capital letters to get a basic idea of how different types of proper nouns were used in the text. We also used regular expressions to help identify and extract individual person names and organizations with common or predictable spelling variations (for instance, *Chrusjtjov, Chrustjof, Chrustjov, Krusjtjev* and *Krustjev*). Likewise, place names were extracted from the resulting list and geotagged through an automated process using the Nominatim API from Open Street Maps[2].

Lemma searches for *terror*, *terrorism* and *terrorist* were performed and the results graphed as trendlines, using both raw frequency and normalized frequency per one million tokens. Normalized lemma frequency was calculated for each individual article by dividing the raw frequency by total number of word tokens. The average normalized frequency for each year was calculated by averaging the normalized frequency of each article in the corpus published during that year. Trendline graphs were created using Microsoft Excel for both single years and years split into 5-year periods. The most frequent compounds for *terror*, *terrorism*, and *terrorist* were extracted, ranked by frequency, and then analyzed.

4.3. The Article Level – Authors & Titles

There are 189 signed and unsigned articles published from 1945–1991 in SvT containing terror-related terminology. Specifically, we are interested in the changing meaning of terror and terrorism and how use of *terror* is different from *terrorism*. We investigated the emergence of the more modern use of *terrorism* as a term that began to refer specifically to sub-state terrorism and non-state actors. This more modern use of *terrorism* began in the 1970s, becoming more frequent later during the Cold War period. When examining articles using *terror* versus *terrorism,* we found a clear dominance of *terror*. *Terror* was used ranging from 1–27 times in 125 articles (on average in 2.7 articles annually) compared to 1–5 times in 27 articles (0.6 articles annually) for *terrorism*. Figure 2 shows this clear dominance of articles with only *terror* simplexes and compounds (red) and also from 1975 (new) *terrorism* (blue and purple) gains traction and increasingly shares the space with *terror*. The graph also includes the 21 articles (in gray) that only contain *terrorist* simplexes and compounds.

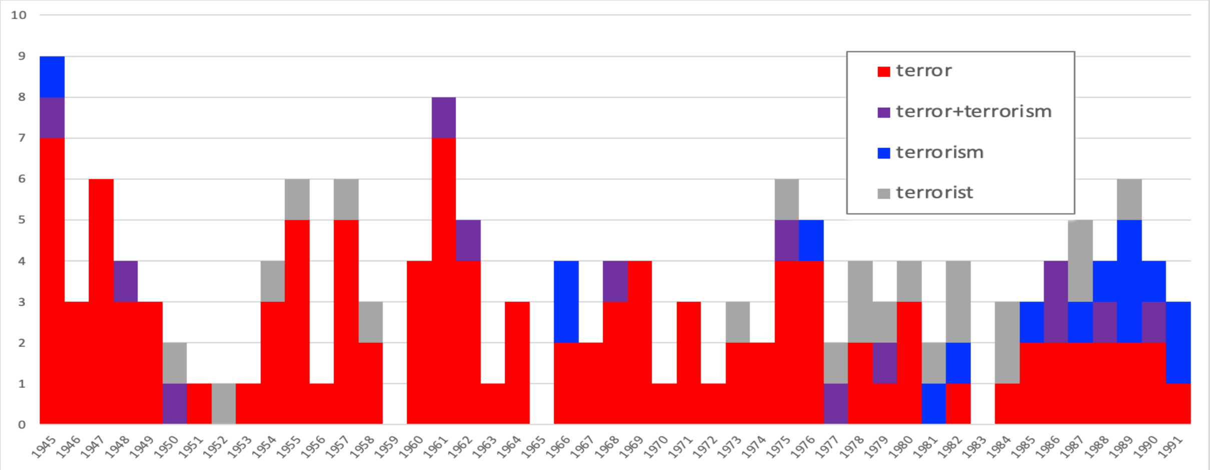


Figure 2: Division of 189 articles containing *terror*, *terrorism* and *terrorist*. Articles containing simplexes and compounds with *terror* (red), with both *terror* and *terrorism (purple), terrorism* only (blue), and *terrorist* only (gray). Black frames mark the years containing the 9 terror-titled articles.

[2] https://wiki.openstreetmap.org/wiki/Nominatim

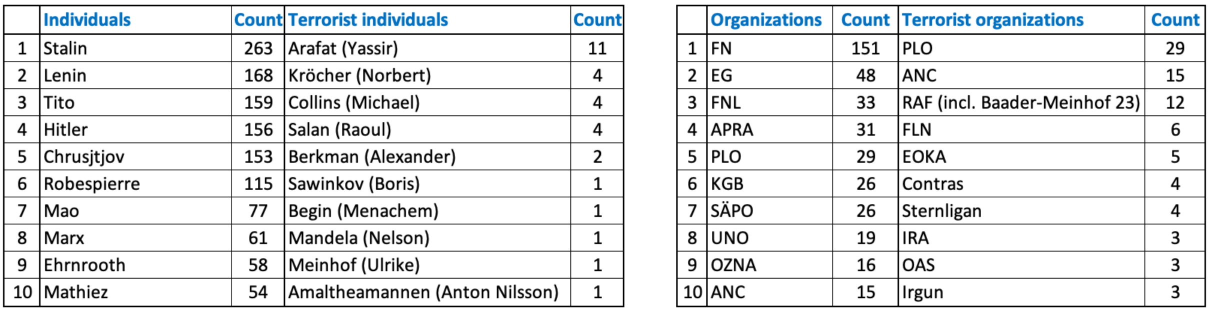
9 articles containing occurrences of *terror* are the most prominent where *terror* or *terrorism* is used in the title in addition to article text (marked with black frames in Figure 2). They are: '*Terrorn och den totalitära staten*' (1946), '*Terrorn – bolsjevikernas politiska instrument*' (1960), '*Om verbal terror*' (1974), '*Terroristerna*' (1975), '*Den revolutionära terrorn*' (1975), '*Om terrorismen*' (1977), '*Statsterrorister och biståndsbanditer*' (1986), '*Stöd ej ANC:s terrorpolitik*' (1987) and '*Den lagliga terrorismen*' (1990). The first two articles contain descriptions of state-related terror in Nazi Germany and the Soviet Union and the next, 'On verbal terror', more than a decade later on metaphorical rather than more literal use describing real-world terror. The new political terrorism of the 1970s is discussed in 3 articles 1975–1977 followed by two critiques of Swedish government policy in the 1980s, discussing how policy relates to terrorism in third world countries. In the last article, 'The legal terrorism' (1990), terrorism is again used metaphorically, to critique the Swedish tax system. Overall, the results reflect contrasting, evolving, and expected use of *terror* and *terrorism*, relating well to what we know about Swedish and Cold War contextual experiences of terror and terrorism.

The SvT corpus contains 84 authors who wrote 115 signed articles. The authors include several well-known Swedish academics and public intellectuals (such as Kristian Gerner, Sven Stolpe and Anders Åslund), politicians (Anders Björck and Jarl Hjalmarsson) and several less known or unknown authors (to be further studied, especially if they are in the project's other cultural periodicals or other materials). From our preliminary research we know some of these writers will be significant in other project source materials, such as the journalist Janerik Larsson who wrote two of the earliest non-fiction books on Swedish terrorism (*Politisk terror i Sverige* ('Political terror in Sweden'), 1968; *Ustasja* 1972) and two terrorist political thrillers (*Attentatet* ('The Attack'), 1975; *Massakern* ('The Massacre'), 1976); the publisher Bo Cavefors who published Swedish translations of political texts of members of the West German Red Army Faction (*Ulrike Meinhofs förbjudna tänkesätt* ('Ulrike Meinhof's Forbidden Thoughts'), 1976; *RAF: texter* ('RAF: Texts'), 1977) that generated debate in the mainstream press; and Thede Palm, member of the SvT editorial board (1967–79) and a former director (1947–65) of the military intelligence agency known as 'the T-office' (*T-kontoret*), and according to one of SvT's chief editors, Palm's 'knowledge and judgment' was 'not infrequently' the 'solid foundation for the journal's positioning within "cultural politics' [86].

A quarter (21) of the authors were 'recurring authors' who each published two or more articles. The top recurring writers are Erik Anners (5 articles in addition to unsigned editorials as chief editor), Tadeusz Norwid (4), Bo Cavefors (3), Birger Hagård (3) Tommy Hansson (3) and Sven Stolpe (3).

4.4. Named Entities - Persons & Organizations

The compilation of the list of the most frequently occurring names of individuals and organizations in SvT terror articles has already shown the potential as well as problems with our distant reading approach.



	Individuals	Count	Terrorist individuals	Count
1	Stalin	263	Arafat (Yassir)	11
2	Lenin	168	Kröcher (Norbert)	4
3	Tito	159	Collins (Michael)	4
4	Hitler	156	Salan (Raoul)	4
5	Chrusjtjov	153	Berkman (Alexander)	2
6	Robespierre	115	Sawinkov (Boris)	1
7	Mao	77	Begin (Menachem)	1
8	Marx	61	Mandela (Nelson)	1
9	Ehrnrooth	58	Meinhof (Ulrike)	1
10	Mathiez	54	Amaltheamannen (Anton Nilsson)	1

	Organizations	Count	Terrorist organizations	Count
1	FN	151	PLO	29
2	EG	48	ANC	15
3	FNL	33	RAF (incl. Baader-Meinhof 23)	12
4	APRA	31	FLN	6
5	PLO	29	EOKA	5
6	KGB	26	Contras	4
7	SÄPO	26	Sternligan	4
8	UNO	19	IRA	3
9	OZNA	16	OAS	3
10	ANC	15	Irgun	3

Table 1 : Most frequently occurring terror-related individual (left) and organization names (right).

In Table 1, a distinction is made between names of contemporary or historical individuals most often associated with political terror in the corpus. The leftmost column lists the overall most frequently mentioned individuals in the corpus primarily connected with terror. Those are followed by 'terrorist individuals' (in the right column) that are individuals occurring in the corpus known to often be – correctly or not – referred to as 'terrorists' in the Cold War debate, with the meaning of being perpetrators of sub-state terrorism. The former 'Individuals' are mostly associated with state terror and terrorism in the original meaning of the word, while the latter 'terrorist individuals' are generally associated with the modern concept of terrorism as conducted by non-state militants. Of course, classifying individuals as 'terrorists' or not is, to a significant extent, a question of perspective, with different classifications and associations expected, depending on the source material. Table 1 shows how frequently certain individuals were associated with 'terror' and 'terrorism'. At a later stage, we intend to perform a more detailed analysis of the context of the word usage. In the case of the category 'terrorist individual', there is also an issue of low frequency, meaning the associations may be less likely to be representative of actual use. Nevertheless, the limited frequency compared to individuals associated with state terror reflects the dominance of state terror/ism in the corpus.

Likewise, the organizations listed as 'Terrorist organizations' in Table 1 are sub-state organizations that are included as often being – correctly or not – designated as such by authorities and media during the Cold War. The most frequently mentioned identifiable organizations do not include national governments due to difficulties of identification but could be included in a later analysis. For example, metonymy makes referent identification more difficult, i.e., Washington can refer to the United States government. However, notably, national security organizations are frequently mentioned in the context of terror and terrorism, three of the top 10 organizations being KGB (Soviet Union), SÄPO (Sweden) and OZNA (Yugoslavia). Notably, CIA (USA) or MI5 and MI6 (UK) are not mentioned at all in the corpus, possibly indicating a conservative and/or Western bias of SvT in its conception of terror and terrorism in this context.

Our distant reading of person names provides a broad and useful understanding of the material, but also reveals methodological limitations of the approach. For example, Ehrnrooth sometimes refers to the Finnish General Casimir Ehrnrooth but also to Adelaide, Gustaf and (the SvT contributing writer) Leo Ehrnrooth, Generally, only the surname is explicitly mentioned, making name referent identification more difficult to determine. Another consideration is the distribution of occurrences of a name. For example, Robespierre is used 115 times, but only in two articles, of which one contains 114 of the references. The French historian (Albert) Mathiez is among the top 10 names but occurs in only one article and is unsurprisingly linked very strongly with Robespierre, *socialism*, and *radikalism*. In our future analysis, we plan to focus both on distribution frequency of names as well as better referent identification. A more thorough approach would help to better contextualize use and association of names.

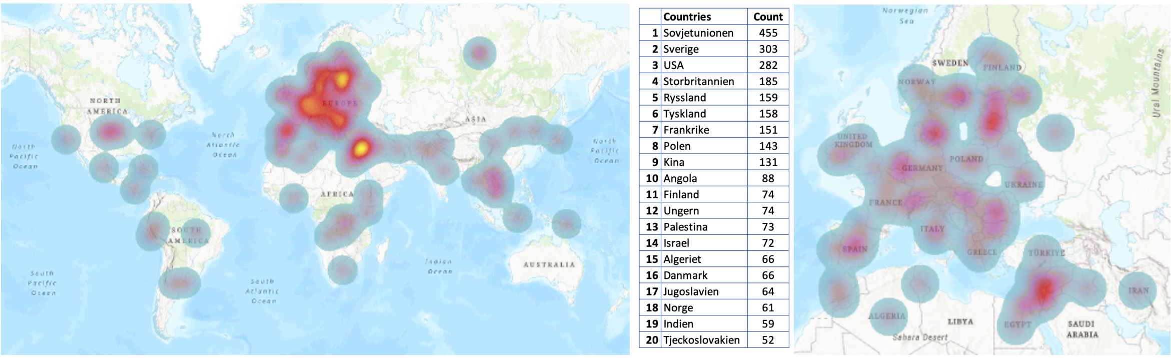
4.5. Named Entities - Locations & Countries

The heat maps in Figure 3 show concentrations of place names, cities and countries, mentioned in the corpus with a minimum cutoff frequency of 5. Maps were created using ArcGIS Online[3]. When it comes to continents, Europe has the densest concentrations with a noticeable hot spot in the Baltics. The Soviet Union and the United States are also prominent. Moreover, Figure 3 reflects how African locations, especially Angola and Congo are frequently mentioned in SvT in relation to decolonial violence and terror. As expected, there is also a hotspot in the Middle East due to interest devoted to the Palestinian conflict and in Southeast Asia because of discussion of the Vietnam War.

While useful for a general overview, the place name extraction and geocoding can also produce some misleading results. While most place names were properly mapped to the correctly named entity, for example *Ryssland* (Russia) and *Sovjetunionen* (Soviet Union) or *USA* and *Amerika* (America), there are limitations. For instance, in some articles Ryssland could refer to Russia within the Soviet Union and in others to the pre-1917 Russian Empire. Likewise, *Tyskland* (Germany) could refer both to the sovereign state after 1991 or its various historical predecessors. While *Amerika* generally refers to the USA, in some contexts, it could refer to North America or South America or (both) the Americas. There are also considerations of how to map a referent to an entire country instead of a more exact city location. For simplicity, we opted for mapping country references to a location near the geographic center of the country. However, for larger countries, this approach could Many references to cities also do not refer

[3] Maps were created using ArcGIS® software by Esri. ArcGIS® and ArcMap™ are the intellectual property of Esri and are used herein under license.

to literal geographic locations. For example, *Washington* and *Moskva* (Moscow) often involve metonymy with the city name referring to the government of the USA or the Soviet Union, respectively. We will pursue these distinctions in more detail in the future but do not see these limitations are problematic in the context of the distant reading approach used in this initial pilot study.



	Countries	Count
1	Sovjetunionen	455
2	Sverige	303
3	USA	282
4	Storbritannien	185
5	Ryssland	159
6	Tyskland	158
7	Frankrike	151
8	Polen	143
9	Kina	131
10	Angola	88
11	Finland	74
12	Ungern	74
13	Palestina	73
14	Israel	72
15	Algeriet	66
16	Danmark	66
17	Jugoslavien	64
18	Norge	61
19	Indien	59
20	Tjeckoslovakien	52

Figure 3: Terror locations. Global and European heat map of place names and list of top-20 countries. Maps created using ArcGIS Online.[4]

4.6. Diachrony - Frequencies of Terror-related Articles and Words

Next, we complemented our analysis of articles containing *terror* and *terrorism* with an analysis involving occurrences of the words. We examined both the raw number of occurrences of tokens *terror* versus *terrorism* in absolute numbers as well as normalized frequency and distribution. When examining overall frequency, *terror* occurs roughly seven times more frequently in simplexes and compounds than *terrorism* (330 to 46 tokens).

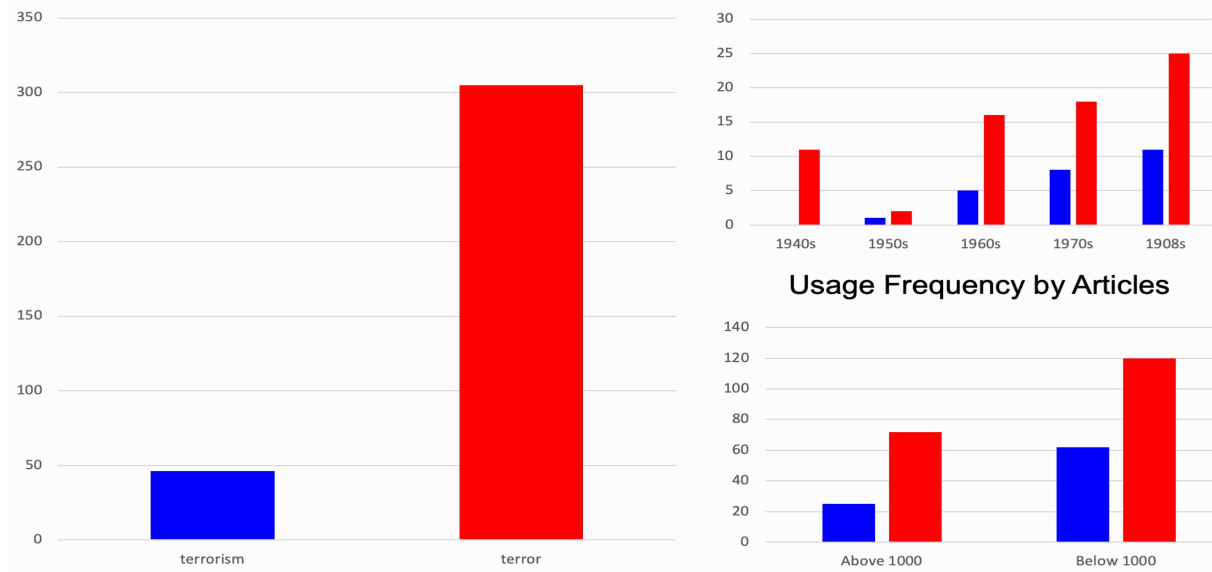


Figure 4: Terror vs. Terrorism and Usage Frequency.

An important distinction is whether the use of terror-words is central to the argument of the article or just spurious, something which we estimated according to a cutoff, classifying articles differently depending on whether they had a normalized occurrence frequency of *terror* or *terrorism* of above or below 1000 per million tokens. The graph on the top right shows articles containing *terrorism* by decade with a normalized frequency below 1000 per million tokens (red) and above 1000 per million tokens

[4] Maps were created using ArcGIS® software by Esri. ArcGIS® and ArcMap™ are the intellectual property of Esri and are used herein under license.

(blue). *Terrorism* occurs in 11 articles with low frequency in the 1940s, reflecting use of *terrorism* with the meaning of causing terror, generally referring to state actors, like Nazi Germany, causing terror during World War II. The use of the word *terrorism* declines sharply in the 1950s and then increases in frequency from the 1970s and onwards as the modern notion of *terrorism*, used to describe non-state actors, develops [4, 5].

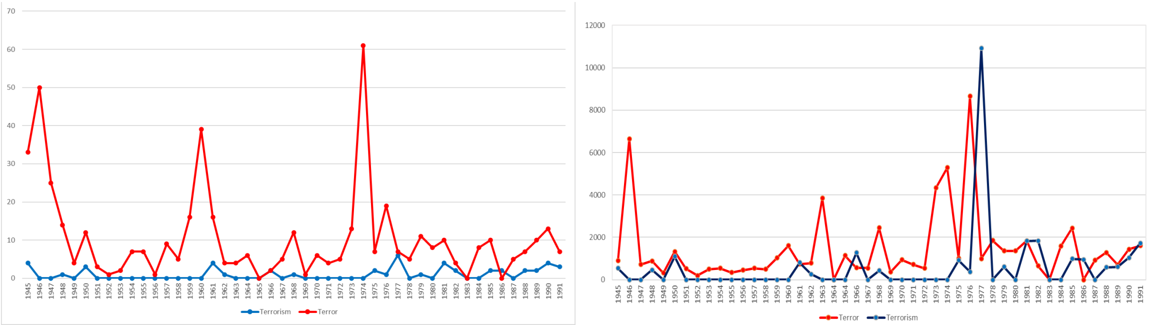


Figure 5: Terror and Terrorism: Total Token (left) and Normalized Frequency (right) Over Time.

The two graphs above show frequencies of *terror* (red) and *terrorism* (blue) over time. The left graph shows absolute token frequency while the right graph shows normalized frequency per one million tokens. Normalized frequency can often be a better measurement as it accounts for differences in article length and variation in the number of total word tokens per year. For example, an absolute frequency can be higher for some years where more terror related articles were published. Some articles are also longer, containing more total tokens. For longer articles, absolute token frequency would be expected to be higher, even if the relative frequency is the same.

Consistent with the previous results, *terror* occurs with a much higher frequency than *terrorism*. The initial peak of *terror* in the 1940s is associated with discussion of events related to World War II, followed by a steep decline in the 1950s. Another increase occurs in the late 1950s and early 1960s, associated with the Cuban Revolution and events related to decolonization in Africa. The strongest peak in the 1970s is associated with the emergence of the modern concept of *terrorism* referring to sub-state and non-state actors. Notably, *terrorism* has a higher normalized frequency than *terror* in the year 1977. This strong spike is associated with one specific article (with 6 mentionings of *terrorism*) discussing West German terrorism in relation to the plot to kidnap a former Swedish Minster.

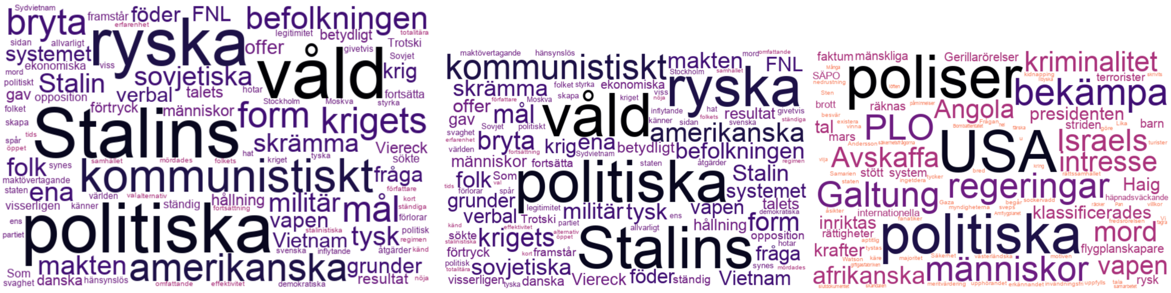


Figure 6: 100 most common words in KWIC (50 words window) of '*terror*' (562 hits), 'terror' (170) and 'terrorism' (26).

We used AntConc with a modified Swedish stoplist (from https://github.com/peterdalle/svensktext/blob/master/stoppord/stoppord-mycket.csv) to produce word clouds of the KWIC (Key Word in Context) window of 25 words before and after the terror-related key words of *terror* and *terrorism,* respectively (Figure 6). These word clouds help contextualize t*error* and *terrorism*. For *terror*, significant contextual words concern the Soviet Union (such as *Stalin/s, ryska, sovjetiska, kommunistiskt, Moskva*) as well as state terror (*makten* ('the power')*, befolkningen* ('the population')*, militär* ('military')*, systemet* (the system')*, förtryck* ('oppression')). For *terrorism*, contextual words are instead more related to international conflicts involving the USA and non-European countries (*PLO, Israels* ('Israel's')*, Angola, afrikanska* ('African')) and combating international and domestic non-state militants and criminals (*poliser* (‘police’), *bekämpa* (‘combat’), *kriminalitet* (‘criminality’)*, mord* (‘murder’)*, SÄPO, flygplanskapare* (‘skyjacker’)*, gerillarörelser* (‘guerilla movements’)). Whether these characterizations will hold for other Swedish cultural periodicals remains to be seen and will be highly interesting to future study.

We also investigated one of the central ideological tensions of the Cold War. Figure 7 shows a trendline of normalized usage frequency for *kapitalism* (‘capitalism’) (blue) and *kommunism* (‘communism’) (red). Communism is often mentioned in the context of terror, terrorism, and the Cold War, with capitalism being mentioned as a positive contrast more or less during the whole period. However, the frequency of the related term *socialism* shows a strong increase in the 1970s. One possibility is that writers in SvT often associated terrorism with socialism in order to mount an ideological critique.

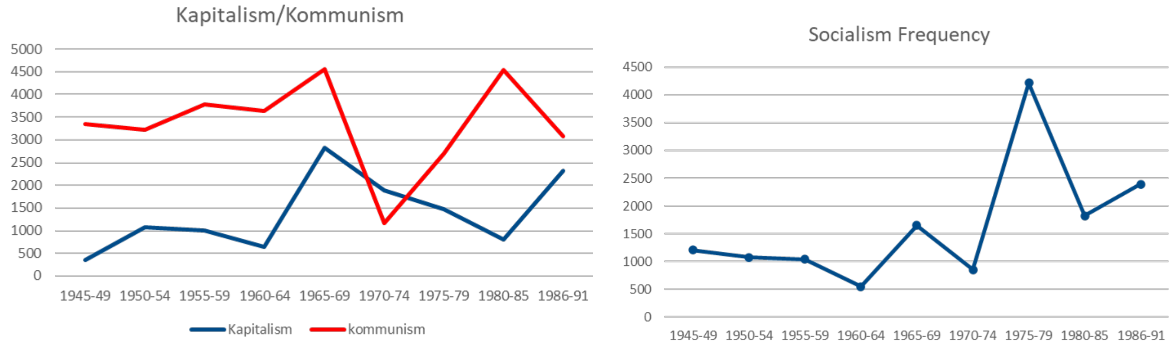


Figure 7: Normalized Frequency of *Kapitalism* (blue) vs *Kommunism* (red) and Socialism (right).

4.7. Linguistics - Innovation and Productivity

Finally, we examined the linguistic innovation and productivity of different compounds consisting of different terror-related words. The most used lemma is terror, of which there exist 44 different compounds with terror as a compound modifier or head. The 10 most common terror compounds are *terrorapparat* (‘terror apparatus’) (16 instances), *terrordåd* ((‘terror deed’) 11), *terrorsystem* (‘system of terror’) (8), *terrormetod* (‘terror method’) (6), *terrorregim* (‘terror regime’) (5), *terrorgrupp* (‘terror group’) (5), *terrorhandling* (‘act of terror’) (5), *terrororganisation* (‘terror organization’) (4) and *terroraktion* (‘terror action’) (3). Of these, 4 (of the top 6) are clearly connected to various forms of state terrorism. The most frequent *terror* compounds all use *terror* as an initial modifying noun in compounds with other nouns as morphological heads. Compounds headed with terror exist but are fewer and with much lower frequency, for example *polisterror* (‘police terror’) (5) and *världsterror* (‘world terror’) (2).

Terrorist occurs in 22 compounds, of which 16 are used as the modifying noun, such as in the most common terrorist compounds *terroristgrupp* (‘terrorist group’) (5), *terroristlag* (‘terrorist law’) (4), *terroristliga* (‘terrorist band’) (3), *terroristorganisation* (‘terrorist organization’) (3) and 6 in which it is in the determined element used for specifying different kinds of terrorists, all only occurring once in the material, with the exception of *statsterrorist* (‘state terrorist’) that are used twice: *ambassadterrorist,* (‘embassy terrorist’) *nykterhetsterrorist* (‘sobriety terrorist’)*, psykoterrorist* (‘psycho terrorist’)*, vänsterterrorist* (left-wing terrorist’) and *yrkesterrorist* (‘professional terrorist’). The adjective simplex *terroristisk* (‘terroristic’) is only used in one compound form, *totalitärterroristisk* (‘totalitarian terroristic’). Also, it is rather remarkable that terrorism is used only once in a (metaphorical) compound; *kvalterrorism* (‘anguish terrorism’). That *terrorism* is so less productive than *terror* indicates that terrorism is a much less salient phenomenon overall than terror during the Cold War period 1945–1991.

5. Conclusions

In this paper, we have presented the integrative approach of the digital history project ‘The Cultural Imaginary of Terrorism’, stressing the benefits of pursuing the intersections of computational and interpretative analysis. By describing the project’s core elements, we have clarified how it contributes to a broader understanding of the cultural meaning-making about terrorism in Sweden in all its diversities, polarizations and complexities. Furthermore, we have shown how we integrate distant and close reading: through a pilot study of the periodical *Svensk Tidskrift*. Notably, our text mining analysis allowed us to explore patterns in the data and our historical domain expertise helped us to discern significant aspects in the text material that call for further investigation (issues, names, etc.). In this sense, our indicative results provide important steps in our efforts to investigate how terrorism was imagined and framed in Swedish culture during the Cold War.

6. Acknowledgements

The Culture Imaginary of Terrorism is funded by the Marcus and Amalia Wallenberg Foundation (https://maw.wallenberg.org/en/political-terrorism-swedish-fact-and-fiction-during-cold-war) together with University of Gothenburg and is carried out in cooperation with the National Library of Sweden (KB) and its digital lab, KBLab. The project is placed at the Gothenburg Research Infrastructure in Digital Humanities (GRIDH), at the Department of Literature, History of Ideas, and Religion at the University of Gothenburg.

7. References

1. Rockwell, G, & Sinclair, S (2016): 'Thinking-through the history of computer-assisted text analysis'. In C Crompton, et al., eds., *Doing digital humanities*. Routledge.
2. Bode, K (2017): 'The equivalence of "close" and "distant" reading'. *Modern language quarterly*, *78*(1).
3. Edlund, J et al. (2022): 'A multimodal digital humanities study of terrorism in Swedish politics'. In K Arai, ed., *Intelligent systems and applications: IntelliSys 2021*, Springer.
4. Stampnitzky, L (2013): *Discipling terror*, Cambridge Univ. Press.
5. Zoller, S (2021): *To Deter and punish,* Columbia University Press.
6. Wilson, T (2022): 'What are terrorism studies?'. In D Muro & T Wilson, eds.,*Contemporary Terrorism Studies*, Oxford University Press.
7. Smyth, M B, J Gunning, R Jackson, G Kassimeris & P Robinson (2008): 'Critical terrorism studies – an introduction', *Critical Studies on Terrorism*, *1*(1).
8. Schurman, B (2018): 'Topics in terrorism research', *Critical Studies on Terrorism*, *12*(3).
9. Jackson, R (2011): 'In defence of 'terrorism'', *Behavioral Sciences of Terrorism and Political Aggression*, 3(2).
10. Brulin, R (2015): 'Compartmentalization, contexts of speech and the Israeli origins of the American discourse on ''terrorism''', *Dialectical Anthropology*, 39.
11. Ditrych, O (2011): *A genealogy of terrorism in states' discourse*. Charles University, Prague.
12. Ditrych, O (2014): *Tracing the discourses of terrorism.* Palgrave Macmillan, Basingstoke.
13. Jensen RB (2018): 'The 1904 assassination of governor general Bobrikov', *Terrorism and Political Violence. 30*(5).
14. Fridlund et al. (2022): 'Trawling and trolling for terrorists in the digital Gulf of Bothnia'. In D. Fišer & A. Witt, eds. *CLARIN*. De Gruyter.
15. Ängsal, M P, et al. (2022): 'Linguistic framing of political terror'. In T Erjavec & M Eskevich, eds., *CLARIN Annual Conference Proceedings.*
16. Steen, L (1990): 'Hinke Bergegren', *Arkiv för studier i arbetarrörelsens historia*.
17. Lång, H (2007): *Drömmen om det ouppnåeliga*, Umeå University.
18. Larsson, P ed. (2008): *Ekot från Amalthea*, ABF Malmö.
19. Lööw, H (1998): *Nazismen i Sverige 1980–1999*, Ordfront.
20. Lööw, H (2004): *Nazismen i Sverige 1924–1979*, Ordfront.
21. Oredsson, S (2001): *Svensk rädsla*, Nordic Academic Press.
22. Oredsson, S (2003): *Svensk oro*, Nordic Academic Press.
23. Flyghed, J, ed. (2000): *Brottsbekämpning*, Studentlitteratur.

24. Flyghed, J & M Hörnqvist, eds. (2003): *Laglöst land,* Ordfront.
25. Bjereld, U & M Demker (2006): *Främlingskap*, Nordic Academic Press.
26. Hansén, D (2007): *Crisis and perspectives on policy change.* Stockholm: Försvarshögskolan.
27. Ekengren, AM, & C Malmström (2004): *Uppvaknandet – Sverige och den internationella terrorismen.* Report. University of Gothenburg.
28. Stenfeldt, J (2013): *Dystopiernas seger*, diss, Lund University.
29. Fridlund, M (2012): 'Det nya gränslösa våldet'. In B Svensson & A Wallete, eds., *Individer i rörelse*. Makadam.
30. Brodén (2008): *Folkhemmets skuggbilder*, diss. Ekholm & Tegebjer
31. Salomon, K, L Larsson & H Arvidsson, eds. (2004): *Hotad idyll*. Nordic Academic Press.
32. Tolonen, M et al. (2018): 'In between research cultures', *Informaatiotutkimus, 37*(2).
33. Marjanen, J et al. (2019): 'A national public sphere?' *Journal of European periodical studies*, *4*(1).
34. Lahti, L et al. (2019): 'Bibliographic data science and the history of the book (c. 1500–1800)', *Cataloging & classification quarterly*, *57*(1).
35. Fridlund, M, et al., eds. (2020): *Digital histories.* Helsinki University Press.
36. Salmi, H et al. (2021): 'The reuse of texts in Finnish newspapers and journals, 1771–1920', *Historical Methods*, *54*(1).
37. Hammar, A N (2015): 'Digital history', *Scandia, 81*(2).
38. Nygren, T et al. (2016): 'Connecting with the past'. In M Hayler & G Griffin, eds., *Research methods for creating and curating data in the digital humanities*, Edinburgh Univ. Press.
39. Orrje, J (2021): 'Vad är digital historia?', *Historisk tidskrift*, *141*(4).
40. Snickars, P (2022): 'Inledning: Digital historia-än sen då?', *Lychnos.*
41. Moretti, F (2005): *Graphs, maps, trees.* Verso.
42. Moretti, F (2013): *Distant reading.* Verso.
43. Jockers, M (2013): *Macroanalysis*, Univ. of Illinois Press.
44. Fridlund, M (2020): 'Digital history 1.5'. In M Fridlund, et al. eds., *Digital histories,* Helsinki Univ. Press.
45. Frank, M (2017): *The cultural imaginary of terrorism in public discourse, literature and film*, Routledge.
46. Sartre, J-P (1940): *L'Imaginare*, Gallimard.
47. Castoriadis, C (1975): *L'Institution imaginaire de la société*, Seuil.
48. Anderson, B (1983): *Imagined communities,* Verso.
49. Ivy, M (1995): *Discourses of the vanishing,* Univ. of Chicago Press.
50. Taylor, C (2004): *Modern social imaginaries*, Duke Univ. Press.
51. Zulaika, J & W Douglass (1996): *Terror and taboo*, Routledge.
52. Strauss, C (2006): 'The imaginary', *Anthropological theory*, *6*(3).
53. Dawson, G (1994): *Soldier heroes*, Routledge.
54. Fluck, W. (1997): *Das kulturelle Imaginäre*, Suhrkamp.
55. Fridlund, M (2011): 'Bollards, buckets and bombs', *History and technology*, *27*(4).
56. Katz, C (2007): 'Banal terrorism'. In D Gregory & A Pred, eds., *Violent geographies*, Routledge.
57. Rychnovska, D (2014): 'Securitization and power of threat framing', *Perspectives*, *22*(2).
58. Azar, M (2001): *Frihet, jämlikhet, brodermord*, Symposion.
59. Azar, M (2006): *Den koloniala bumerangen*, Symposion.
60. Balibar, É & I Wallerstein (1988): *Race, nation, class*, Verso.
61. Spivak, G C (1993): *Outside in the teaching machine,* Routledge.
62. Rajan, G, & Mohanram, R. eds. (1995): *Postcolonial discourse and changing cultural contexts.* Greenwood Publishing Group.
63. Nord, L (2001): *Vår tids ledare.* Carlssons.
64. Ahmad, J (2018): *The BBC, the war on terror and the discourse construction of terrorism*. Palgrave Macmillan.
65. Roosvall, A (2022): 'Kulturjournalistikens politiska dimensioner'. In K. Riegert, et al. eds. *Kulturjournalistikens världar*, Nordic Academic Press.
66. Appelbaum, R, Paknadel, A (2008): 'Terrorism and the novel, 1970–2001', *Poetics today, 29*(3).
67. Herman, P C, ed. (2018): *Terrorism and literature*, Cambridge Univ. Press.

68. Yao, Z, et al. (2017): ‘Dynamic word embeddings for evolving semantic discovery’, *International conference on web search and data mining,* WSDM.
69. Underwood, T (2019): *Distant Horizons*, Univ. of Chicago Press.
70. Berry, D & A Fagerjord (2017): *Digital humanities*, Polity.
71. Dobson, J E (2019): *Critical Digital Humanities*. Univ. of Illinois Press.
72. Leifeld, P (2017): ‘Discourse Network Analysis'. In J.N. Victor, et al. eds., *The Oxford handbook of political networks*, Oxford University Press.
73. Venturini, T, & A. K. Munk (2021): *Controversy mapping*, John Wiley & Sons.
74. Callon, M (1990): ‘Techno-economic networks & irreversibility’, *Sociological review*, *38*(S1).
75. Latour, B (1999): *Pandora's hope*, Harvard Univ. Press.
76. Latour, B (2007): *Reassembling the Social*, Oxford Univ. Press.
77. SvT (1963): ‘Svensk Tidskrift i fransk edition’, *Svensk Tidskrift*, *49*.
78. Anner. E (1962): ’Angolakrisen inför svensk opinion’, *Svensk Tidskrift*, *48*.
79. Johansson, M (2011): ‘En publicistisk gärning'. In M. Johansson, ed. *Texter i tiden,* Timbro.
80. Schmid, H (1995): ‘Improvements in Part-of-Speech Tagging with an Application to German’, *Proceedings of the ACL SIGDAT-workshop*.
81. Schmid, H (1994): ‘Probabilistic part-of-speech tagging using decision trees’, *Proceedings of International Conference on New Methods in Language Processing*, Manchester, UK.
82. TXM Team (2013):' *TXM Manual*. ICAR Laboratory, Lyon University & CNRS, Lyon, France. https://txm.gitpages.huma-num.fr/textometrie?lang=en
83. Heiden, S (2010): ‘The TXM Platform’. In Otoguro R, Ishikawa K (eds.), *Proceedings of the 24th Pacific Asia Conference on Language, Information and Computation (PACLIC 24). 4-7 November 2010, Sendai*.
84. Lafon, P (1980): ‘Sur la variabilité de la fréquence des formes dans un corpus’, *Mots*, *1*(1).
85. Anthony, L (2022): AntConc (Version 4.2.0) [Computer Software]. Tokyo, Japan: Waseda University. Available at https://www.laurenceanthony.net/software
86. Anners, E (1996): ’Till minnet av Thede Palm’, *Svensk Tidskrift, 82*.

A Complex Philosophical Oeuvre and Its Complex User Community: Reflections on the Past, Present, and Future Digitisation of Wittgenstein's Philosophical Writings

Nivedita Gangopadhyay*[1], Sebastian Sunday Grève*[2,3] and Alois Pichler[4]

[1] *University of Bergen*
[2] *Chinese Institute of Foreign Philosophy*
[3] *Peking University*
[4] *University of Bergen*

Abstract

This paper reports the results of a user survey conducted by the Wittgenstein Archives at the University of Bergen (WAB) concerning some of the digital research tools and resources it has developed over the past three decades. The authors' analysis of the survey results is embedded within a broader discussion of the nature and history of the digitisation of Wittgenstein's philosophical writings. It is argued that the special nature of both Wittgenstein's philosophy—especially the later Wittgenstein's conceptions of meaning, concepts, and philosophy itself—and its primary material sources (Wittgenstein's "Nachlass") inherently call for a variety of advanced digital tools and resources; for instance, a digital *interactive* edition of his works, such as WAB's Wittgenstein IDP: The Nachlass in Interactive Dynamic Presentation, to complement static ones such as traditional print editions and other, non-interactive digital editions. For similar reasons, the authors argue, WAB's ongoing and future development of digital tools and resources for the study of Wittgenstein's philosophy must be closely coordinated with the evolving needs of WAB's large and diverse user community.

Keywords

Digital humanities, philosophy, Wittgenstein Archives, interactive dynamic scholarly edition, research platform, semantic technology, ontology

1. Introduction

In this paper, we discuss the results and implications of a user survey we have conducted regarding the digital tools and resources provided by the Wittgenstein Archives at the University of Bergen (WAB). We argue that the nature of Wittgenstein's philosophical oeuvre strongly requires the development of advanced digital tools and resources to serve as user interfaces, and that this development must be closely coordinated with users' evolving needs.

DHNB2023 | Sustainability: Environment – Community – Data. The 7th Digital Humanities in the Nordic and Baltic Countries Conference. Oslo – Stavanger – Bergen, Norway. 8–10 March 2023.

*Co-first authors

nivedita.gangopadhyay@uib.no (N. Gangopadhyay*); sebastian.sunday.greve@gmail.com (S. Sunday Grève*); alois.pichler@uib.no (A. Pichler)

DHNB Publications, DHNB2023 Conference Proceedings, https://journals.uio.no/dhnbpub/issue/view/875

WAB was established in 1990. It is a research infrastructure and project platform that brings together philosophy, editorial philology, text technology, and digital humanities. Its holdings comprise works by the philosopher Ludwig Wittgenstein (1889–1951). WAB is perhaps best known for the publication of *Wittgenstein's Nachlass: The Bergen Electronic Edition* (Oxford University Press, 2000). Wittgenstein left behind around 20,000 pages of unpublished writings, his "Nachlass". WAB's research infrastructure includes digital and paper copies as well as transcriptions of the Nachlass, following Georg Henrik von Wright's catalogue system from 1969 (see Wright 1982 and the reprint with an addendum in Wittgenstein 1993). Since the publication of the *Bergen Electronic Edition*, WAB's digital tools and resources have undergone substantial further development. In particular, since 2012 WAB has offered Semantic Faceted Search and Browsing (SFB), and since 2014 it has produced new digital facsimiles of the Wittgenstein Nachlass, which are available open-access via Wittgenstein Source. Since 2016, WAB has enabled interactive open access to all its transcriptions of the Wittgenstein Nachlass using an Interactive Dynamic Presentation (IDP) tool. In addition, WAB has developed an advanced Nachlass search tool in cooperation with the Center for Information and Language Processing at LMU Munich: the FinderApp WiTTFind. Today, these and other resources are all easily accessible via the newly created portal wittgensteinonline.no.

WAB has an extensive and diverse user base, comprising researchers from disciplines such as philosophy, computational linguistics, digital humanities, philology, literary theory and criticism, art, graphic design, and musicology. Recently, we launched a user survey about WAB's main digital tools and the resources available via wittgensteinonline.no. The survey asked users to evaluate the following digital tools and resources:

Wittgenstein IDP	The Nachlass in Interactive Dynamic Presentation
Wittgenstein SFB	Wittgenstein Resources by Semantic Faceted Search and Browsing
Wittgenstein Source	The Bergen Nachlass Edition and Other Primary Sources
WiTTFind	The FinderApp for Nachlass Text Search
Wittgenstein XML TEI	The Nachlass in XML TEI Transcription
Wittgenstein OWL	Wittgenstein Resources in Ontology Representation

The target user group for the initial phase of the survey was selected on the basis of long-term participation in the Wittgenstein research community and familiarity with WAB's digital resources. The main goal of the user survey is to strengthen user-oriented development and to integrate users' feedback into decision making processes.

Wittgenstein's Nachlass is enormously complex, in terms of its philosophical content, the diversity of materials it contains, the open-ended uses to which it might be put, and the sheer number of manuscript pages. Consequently, users' participation in developing WAB's digital resources seems vital, and this is something that has long been taken very seriously at WAB. Wittgenstein's works represent a classic case of a humanities oeuvre that resists a generalised, mechanical structuring that lacks sensitivity to its subject-specific (in this case, philosophical) subtleties. As such, one can think of WAB as a constant work in progress, much like Wittgenstein envisaged his philosophy to be. Thus, the success of WAB as a knowledge base and research tool critically depends on its being appropriately dynamic, just like the works it represents and the kinds of use it is intended to enable; and so user involvement in the development and maintenance of the resources is key to achieving this goal. In light of this, it is perhaps unsurprising that one piece of positive feedback that stood out in our user survey is that users especially appreciate the interactive features of certain resources, which allow them to adopt something like an editorial role for themselves.

In this paper, we shall first briefly introduce Wittgenstein's life and works (section 2) and outline WAB's project of digitising Wittgenstein's works and some of the inherent difficulties it involves (section 3), before discussing some of the results of our survey and their implications for the future development of WAB's digital resources (section 4).

2. Wittgenstein's life and works

The philosopher Ludwig Wittgenstein (1889–1951) is widely considered one of the greatest philosophers of the 20th century. His work was exceptionally influential in the Anglo-American tradition, revolutionising what is known as analytic philosophy not once but twice, first with his *Tractatus Logico-Philosophicus* (1922) and then with the posthumously published *Philosophical Investigations* (1953/2009). Wittgenstein's life and philosophical thoughts stand out for their uniqueness and authenticity. He was a brilliant, complicated thinker who refused to conform with traditional ways of living and philosophising.

Wittgenstein was born in Vienna on 26 April 1889 into one of the wealthiest families of the Austro-Hungarian Empire. As a boy he was deeply interested in engineering. In 1908, he went to Manchester to study aeronautics. But he became increasingly interested in the problems of pure mathematics, and in 1911 he travelled to Cambridge to meet Bertrand Russell, author of the monumental *Principia Mathematica* (with A. W. Whitehead; first published 1910). After meeting Russell, Wittgenstein abandoned his studies of aeronautics and devoted himself to logic and philosophy. At the outbreak of WW1 in 1914, he enlisted as a soldier in the Austrian army. He fought in many battles and was awarded several medals for valour. And yet, by the end of the war, he had managed to complete his first book, the *Tractatus*, which would remain the only philosophical book he would publish in his lifetime. He then took a hiatus from academic philosophy until 1929, when he returned to Cambridge. In Cambridge, he lived and worked at Trinity College supported by a fellowship and teaching work, eventually succeeding G. E. Moore as professor of philosophy in 1939. In 1947, Wittgenstein resigned from his chair at the Faculty of Philosophy and moved to Ireland. He died of cancer in Cambridge in 1951.

Wittgenstein's philosophical work can be divided into two broad categories, the early Wittgenstein and the later Wittgenstein. The early phase of Wittgenstein's philosophical thoughts and writings culminated in his *Tractatus.* The *Tractatus* was first published in German in 1921, but Wittgenstein despised this first edition. A translation into English, by C. K. Ogden (with the help of Frank Ramsey), was published by Kegan Paul in 1922 alongside the original German text. The work was significantly influenced by the intense discussions about logic and the nature of philosophy that Wittgenstein had had with Russell and Moore and the economist J. M. Keynes in Cambridge between 1911 and 1913.

The ambition of the *Tractatus* was to present nothing less than a comprehensive account of philosophy in general. We cannot go into the philosophical intricacies of *Tractatus* exegesis here. However, it is important to note the stark contrast between Wittgenstein's early and later works, because it is as a consequence of these philosophical and exegetical intricacies that a number of important issues arise for any attempt to prepare a digital scholarly edition of such material: that is, the oeuvre of an author whose philosophical thinking underwent a long period of revolution in the form of constant and profound self-criticism that is substantially reflected in his writings. Like most philosophical texts, Wittgenstein's *Tractatus* has been interpreted in many different ways. In fact, the *Tractatus* is widely viewed as perhaps one of the most difficult texts to interpret in the canon of analytic philosophy. Notably, Wittgenstein himself thought little of the introduction that accompanied the 1922 bilingual edition, which was written by his friend and former teacher Russell. Wittgenstein thought Russell had fundamentally misunderstood the book. Later, in 1929, Russell and Moore were more than happy to pass the book as Wittgenstein's PhD dissertation. At the conclusion of the official oral exam, Wittgenstein put an arm around both his examiners' shoulders and said, "Don't worry, I know you'll *never* understand it" (see Wood 1957, 156).

Wittgenstein had periods of intense reclusiveness throughout his lifetime. For example, he would sometimes retreat for weeks on end to an isolated hut he had erected for himself in the remote Skjolden fjord in central Norway, where he would ponder philosophical problems. The reason Wittgenstein withdrew from academic philosophy after the publication of the *Tractatus* was that he believed himself "to have found, on all essential points, the final solution of the problems" (1922, preface). He gave away all his inherited family fortune, and for several years pursued a variety of professions in his native Austria, including as a schoolteacher, gardener, and architect. He was eventually persuaded to return to Cambridge and academic philosophy in 1929, largely thanks to the efforts of British mathematician and philosopher Frank Ramsey (who, as noted above, had previously played a role in the translation of the *Tractatus*).

In many ways, Wittgenstein was his own best critic. For example, one of the fundamental ideas in the *Tractatus* is that philosophy is not a doctrine but an activity, and so cannot be treated dogmatically. But after his return Wittgenstein criticised his earlier work in the *Tractatus* for still being far too dogmatic. Subsequently, his thinking and method developed rapidly in a number of new directions. The culmination of the most radical changes can be found in his posthumously published *Philosophical Investigations* (1953/2009). This is the work generally taken to authoritatively represent the later Wittgenstein.

Between the early and later Wittgensteins, there is sometimes said to be a middle Wittgenstein (see, for example, Stern 1991). This "middle period" is said to have begun shortly after Wittgenstein's return to Cambridge, during which his thought underwent several radical changes. In addition to his notebooks, there are several volumes of conversations, letters and lecture notes that record some of the intense philosophical exchanges he was engaged in during this time, including with colleagues in Cambridge and Vienna. It was then that he decided to circulate what is now known as the *Blue Book* (1958/1969), a dictated typescript that was partly born out of his frustration at being (or, at any rate, feeling) misunderstood by his peers, including in their published secondary accounts of his work. Moreover, many Wittgenstein scholars have argued that there is another, post-*Investigations* Wittgenstein, i.e. an incarnation that is later than the later Wittgenstein, who finally retreats from his anti-dogmatism and actually (and knowingly) advances philosophical theses (for instance, in the well-known posthumous edition published as *On Certainty*, 1969). On the present count, that would be a *fourth* Wittgenstein, but in fact this school of interpretation is commonly known as "the third Wittgenstein" (see, for instance, Moyal-Sharrock 2004). David Stern's article "How many Wittgensteins?" (2006) offers a useful overview of the complicated terrain of contemporary Wittgenstein scholarship.

Upon his death in 1951, Wittgenstein left behind a philosophical Nachlass of some 20,000 pages. In his will, he appointed three of his closest personal and philosophical associates—Rush Rhees, Elizabeth Anscombe, and Georg Henrik von Wright—as his literary trustees, with the instruction to publish from his Nachlass at their own discretion. The resulting body of posthumous publications, consisting of various editions of unpublished philosophical notebooks, manuscripts, typescripts, and dictations and spanning the period of 1913 to 1951, reveals not only Wittgenstein's complex life work but also a fascinating and complex relationship between the work itself (the raw material) and its presentation to the general public, as shaped and coloured by the editors' interpretative practices (for a historical overview, see Erbacher 2020).

3. Digitising Wittgenstein

3.1. The Bergen Electronic Edition

The year 2000 saw the publication of *Wittgenstein's Nachlass: The Bergen Electronic Edition* (BEE), a joint publication by the Wittgenstein Archives at the University of Bergen (WAB) and Oxford University Press. BEE was the result of more than ten years of focused research and editorial work (see esp. Huitfeldt 1994). In addition to containing sources and drafts of some of the most important book editions of Wittgenstein's work, BEE made available a significant quantity of previously unpublished material.

Wittgenstein's manuscripts contain many instances of overwriting, substitutions, and deletions (with the deleted parts occasionally illegible); there are also a significant number of spelling mistakes and passages written in code, as well as the occasional doodle. In addition to facsimiles, BEE makes all of these details available in both "normalised" and "diplomatic" transcribed versions. The diplomatic version is designed to represent the author at work, allowing users to chart the course of Wittgenstein's thought as it progresses, and sometimes digresses, in writing. The normalised version is tidier: for example, deleted text is omitted, spelling is corrected, and decisions are made between alternatives (usually in favour of the variant that was last added). Thus, BEE consists of three sub-editions—the diplomatic version, the normalised version, and the facsimiles—and could therefore also be called a "combined edition" (Pichler and Haugen 2005).

BEE represents the earliest attempt to digitise Wittgenstein's complex philosophical oeuvre. At the time of its publication, it provided Wittgenstein scholarship with an unprecedented, digital form of access to his Nachlass and so opened up new research possibilities (see also Meschini 2020). Simultaneously, it made the Nachlass available to a large readership across the world. Thus, BEE clearly demonstrated some of the advantages of a digital edition over a print one, especially for a body of writings as complex as Wittgenstein's Nachlass, including giving users more flexibility in how they could access it (Pichler 2021). In general, BEE remains an invaluable resource as the record of a pioneering attempt to digitise one of the most complex philosophical oeuvres that exists.

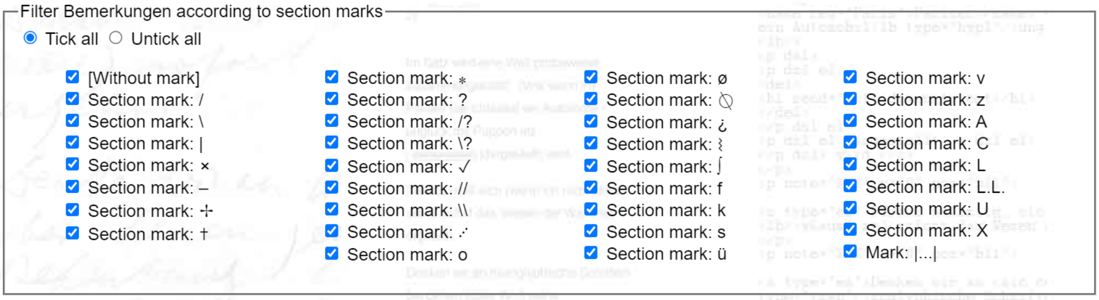
Today, the nature and complexity of Wittgenstein's philosophical oeuvre, paired with the evolving and dynamic needs of the user community and the rapid development of digital technologies, requires a digital humanities platform that offers more flexibility and more active involvement from the user community than is possible with BEE. As crucial as BEE has been to the project of digitising Wittgenstein's Nachlass, it remains a *static* scholarly edition. Therefore, after more than a decade of testing on a subset of the Nachlass—an early pilot was conducted in 2004 (see https://wab.uib.no/sept1914/home.html; early discussion in Hrachovec 2000)—and having obtained all required permissions, in 2016 WAB launched a complete, *interactive* edition called Wittgenstein IDP: The Nachlass in Interactive Dynamic Presentation, which grants open access to all of WAB's transcriptions (available at wittgensteinonline.no).

3.2. Wittgenstein IDP: The Nachlass in Interactive Dynamic Presentation

There is a major difference between the editorial practices of BEE and IDP. In the case of BEE, all editorial decisions were taken by the official editors, in particular regarding the precise level and representation of detail (deletions, variants, section marks, etc.) in the diplomatic version. But in the case of IDP, every user can adopt an active, editorial role for themselves and decide how exactly they want the material to be presented in accordance with their own particular preferences (Pichler and Bruvik 2014). According to our survey (reported in section 4 below), IDP is the second most popular digital tool offered by WAB, after the *Bergen Nachlass Edition* (BNE) on Wittgenstein Source.

The introduction and development of IDP was motivated by WAB's recognition of a growing and deeply rooted need in the user community, which was fuelled by the special character of Wittgenstein's philosophical practice and oeuvre. For instance, the dialogical character of Wittgenstein's philosophical practice and writing, especially in his later works, is partly a function of his view that philosophy is a kind of intellectual therapy. As a consequence, he took it to be his duty as a teacher and writer of philosophy that he and his texts should enable others to work through their own individual problems and confusions. Thus, in Wittgenstein's own view his readers must be, and must be enabled to be, active interrogators rather than passive recipients. This clearly already applies to the early Wittgenstein and the *Tractatus*, but is especially true of the later Wittgenstein (see Sunday Grève 2015). More generally, it follows from this that in order to be maximally faithful to Wittgenstein's approach to philosophy, a scholarly digital edition of his Nachlass should do more than adequately capture all of Wittgenstein's edits and revisions; in particular, it should also offer possibilities that go beyond any static edition.

A useful example in this connection is Wittgenstein's employment of section marks (of which the Nachlass contains more than twenty thousand in total; see also Figure 1). A digital edition that simply represents the relevant symbols on the page, even if placed in their precise location and so on, misses the crucial point that Wittgenstein used these symbols for a reason: he wanted to *do* things with the text sections he marked. Sometimes, he wanted to have a given set of remarks dictated; in other cases, he wanted them to be copied, rearranged, or omitted. Thus, Wittgenstein's section marks express "action intentions" (Pichler 2021, 197).



Filter Bemerkungen according to section marks

● Tick all ○ Untick all

☑ [Without mark]
☑ Section mark: /
☑ Section mark: \
☑ Section mark: |
☑ Section mark: ×
☑ Section mark: –
☑ Section mark: ⸭
☑ Section mark: †
☑ Section mark: ⁎
☑ Section mark: ?
☑ Section mark: /?
☑ Section mark: \?
☑ Section mark: √
☑ Section mark: //
☑ Section mark: \\
☑ Section mark: ⋰
☑ Section mark: o
☑ Section mark: ø
☑ Section mark: ⦰
☑ Section mark: ¿
☑ Section mark: ⦚
☑ Section mark: ∫
☑ Section mark: f
☑ Section mark: k
☑ Section mark: s
☑ Section mark: ü
☑ Section mark: v
☑ Section mark: z
☑ Section mark: A
☑ Section mark: C
☑ Section mark: L
☑ Section mark: L.L.
☑ Section mark: U
☑ Section mark: X
☑ Mark: |...|

Figure 1: Section mark selection menu from Wittgenstein IDP

So rather than merely receiving a visual representation of the symbol, the reader should be able to understand and explore Wittgenstein's section marks as the action intentions that they are; indeed, they should be able to act those intentions out for themselves, for instance by grouping and displaying together remarks with the same section mark. This is precisely the sort of function that IDP was designed to offer.

Moreover, in the case of section marks, the reader's editorial role is especially important for yet another reason. Since the meaning of some of Wittgenstein's symbols is still only partially understood, the reader must be given the freedom to represent the text (arrange or omit sections, etc.) according to their own interpretation. Obviously, in order to provide the reader with this kind of freedom a scholarly digital edition must, at a minimum, provide the reader with the option to do so and, ideally, with the right kinds of tools in order to do so effectively. Again, this is precisely the sort of thing that IDP was designed to offer.

Thus, IDP exemplifies a progressive way of developing a digital tool for a complex philosophical oeuvre that closely follows—that is, both monitors and serves—the needs of the user community, and which no static edition could replicate.

3.3. Wittgenstein SFB, WiTTFind, and OWL

At the same time, the project of developing an adequate and comprehensive digital platform for an oeuvre as complex as Wittgenstein's will clearly need to go further still. Today, such a project must also provide at least a minimal semantic framework. Semantics, in this connection, is the battlefield where digitisation truly runs up against the challenge that lies at the heart of all digital humanities research: the digitisation of meaning. Since 2012, WAB has therefore also provided Wittgenstein SFB: Wittgenstein Resources by Semantic Faceted Search and Browsing, which, just like all other current digital tools provided by WAB, is now also available via the recently created portal wittgensteinonline.no.

The term "faceted" (the "F" in SFB) simply refers to the properties and relations ("facets") in terms of which the domain's objects (i.e. text) can be classified. Thus, in SFB text and metadata are combined. SFB still needs a lot more development (see, for example, Pichler 2021). But it is already a powerful tool for implementing metadata, such as dates of composition, which works or individuals are being referred to, and which published edition of Wittgenstein's works a given remark was published in. However, Wittgenstein's ideas, use of expressions, and spelling changed over the course of his lifetime; a state of affairs that is perhaps to be expected in the case of any complex philosophical oeuvre. As a consequence, digital semantic technology such as SFB requires a lemmatised lexicon. In the case of SFB, this is currently a work in progress, as the WiTTFind lexicon (Röhrer 2019; see also Hadersbeck et al. 2020). is still in the process of being implemented.

For similar reasons, building a computational ontology is crucial for proper digitisation of a complex oeuvre such as Wittgenstein's. WAB currently offers a pilot version of a computational ontology on Wittgenstein OWL: Wittgenstein Resources in Ontology Representation ("OWL" stands for Web Ontology Language).

Wittgenstein's philosophical oeuvre remains one of the most semantically dynamic and interpretatively contested in history. In particular, the Nachlass itself contains a large number of competing knowledge claims. At the same time, however, Wittgenstein's philosophy also suggests possible ways forward in building a suitable ontology. In the *Investigations*, the later Wittgenstein introduces the notion of "crisscross" conceptual structures and methods of enquiry (see Pichler 2016). In the course of developing the notion of a language-game, he insightfully lays bare the difficulties in finding properties that are common to all games, before concluding:

> … And the upshot of these considerations is: we see a complicated network of similarities overlapping and criss-crossing: similarities in the large and in the small.
>
> 67. I can think of no better expression to characterize these similarities than "family resemblances"; for the various resemblances between members of a family – build, features, colour of eyes, gait, temperament, and so on and so forth – overlap and criss-cross in the same way. – And I shall say: 'games' form a family. (1953/2009, secs 66–67; see also the preface)

Wittgenstein argued that not only the concept *game* but many of our most fundamental concepts (*language*, *number*, *meaning*, *knowledge*, etc.) share this crisscross-type structure. This account of concepts has profound implications for the digitisation of Wittgenstein's philosophical writings, and potentially for computational ontology in general. For Wittgenstein's account shows at least that it is not necessary, neither practically nor theoretically, that all entities subsumed under a general term of a natural language must have something essential in common by virtue of which they are so grouped.

Wittgenstein's analysis thus raises the question of how general terms may be thought to work instead. What determines whether a particular general term is applied correctly to a given case or not? What is it that unites the things that are subsumed under a general term such as "game"? Here the notion of language-games comes into play (see Sunday Grève 2018). Wittgenstein writes:

> The word "language-*game*" is used here to emphasize the fact that the *speaking* of language is part of an activity, or of a form of life. (Wittgenstein 1953/2009, sec. 23)

The meanings of words in our natural languages, and with them the structures of the concepts that correspond to them, evolve organically in the dynamic situated contexts and environments in which they are used by the members of a given linguistic community. For example, children learn the meanings of words by learning what to do with words on the basis of examples and on particular occasions. In this way, children are initiated into the linguistic practice of their community, which they inherit. If this is how meaning and conceptual structures are constituted at the level of linguistic tradition, then no one in a given community of speakers of a natural language need be able to formulate the rules according to which the words of the language are to be used; moreover, there need exist no fixed rules that could be formulated (except perhaps in the most abstract way, if we froze time and surveyed all present and past facts about usage); on the contrary, the nature of meaning and concepts will be essentially dynamic, open-ended, and evolving (see Sunday Grève forthcoming). Wittgenstein thought that using language in general resembles the kinds of games that young children are taught to play during the early phase of natural-language acquisition, and so he coined the term "language-game" to refer to virtually all kinds of language use.

Thus, Wittgenstein's account of the nature and structure of meaning and concepts provides a framework within which an appropriate computational ontology may be conceived, especially perhaps for Wittgenstein's own complex oeuvre and others in a similar mould. However, the detailed theoretical conception and practical development of this envisioned framework, not to mention its concrete implementation, remains a major challenge (see Pichler et al. 2021). The pilot currently available on Wittgenstein OWL employs the RDF (Resource Description Framework) data model in an attempt to organise the entire Wittgenstein domain—not only the Nachlass but also other sources (including scholarly secondary literature)—under three top classes: Source, Person, and Subject (Pichler and Zöllner-Weber 2013). However, faithful representation and implementation of Wittgenstein's philosophy in the form of a computational ontology—especially if it is to respect Wittgenstein's own account of crisscross conceptual structures—will require novel approaches in ontology design itself. The first step is to try to render an ontological framework that may be capable of fully representing the continuous evolution of meaning, competing knowledge claims, and multiperspectivism that is characteristic of Wittgenstein's work.

To date, such an ontology remains an unattained ideal, but WAB's pilot is a proof-of-concept project. In particular, the aim is to show how in principle to digitally map a rich body of multicultural humanities knowledge within a computational ontology environment that incorporates competing structures (meanings, concepts, knowledge claims, etc.) into a single model.

4. WAB and its users

Today's challenges of digitising Wittgenstein's philosophical works are not merely a consequence of the material and intellectual nature of the works themselves, but are intimately bound up with the user community too. Of course, certain characteristics of any given user community can be predicted with a fair amount of certainty. However, it should be equally obvious that the user community of a complex philosophical oeuvre such as Wittgenstein's, including associated modern digital resources such as those offered by the Wittgenstein Archives at the University of Bergen (WAB), is itself no less complex an animal, and one which could develop in all kinds of unforeseen directions.

For this reason, we have recently started systematically surveying users about their views on and experiences with WAB's current main digital tools and resources, available via wittgensteinonline.no. WAB's user base is a large and diverse group, including students and scholars of philosophy, computational linguistics, digital humanities, philology, literary theory and criticism, graphic design, and musicology. However, the target users during the initial phase of the survey, results from which we present in this paper, belong to a small group of experienced scholars who were selected on the basis of their long-term participation in the Wittgenstein research community and their high degree of familiarity with WAB's digital resources.

The survey asked users to evaluate the following digital tools and resources, which are available (open-access) via wittgensteinonline.no.

Wittgenstein IDP	The Nachlass in Interactive Dynamic Presentation
Wittgenstein SFB	Wittgenstein Resources by Semantic Faceted Search and Browsing
Wittgenstein Source	The Bergen Nachlass Edition and Other Primary Sources
WiTTFind	The FinderApp for Nachlass Text Search
Wittgenstein XML TEI	The Nachlass in XML TEI Transcription
Wittgenstein OWL	Wittgenstein Resources in Ontology Representation

The survey was conducted using Google Forms. It consisted of the following 18 questions. Questions (15) and (16) were designed to give participants the option to remain anonymous.

1) How did you first hear about WAB resources?
2) Which of the following digital tools are you familiar with or have you used? IDP, SFB, Wittgenstein Source, WiTTFind, Wittgenstein XML TEI, Wittgenstein OWL.
3) Which digital tool do you use the most?
4) Do you easily and usually find what you are looking for when visiting WAB resources?
5) What, if any, are the challenges you face when using WAB resources?
6) Are there any other online Wittgenstein resources, other than WAB, that you use? If yes, please specify.
7) For what research question(s) or research project(s) have you used digital tools or methods? Please specify the digital tool used.
8) Briefly describe the advantage(s) of using a digital tool in your research. Did it allow you to ask different research questions? Did it allow you to complete a research process more quickly or efficiently? Did it help in some other way?
9) Are there any other digital tools that do not yet exist that would be helpful to your research? Please describe the proposed digital tool and how it would help your research.
10) How likely is it that you would recommend WAB resources to colleagues? Please answer with a number between 1 and 10. (10 – extremely likely. 1 – not at all likely)
11) If you recommend WAB resources to your colleagues, which of the following would you recommend? Please choose from the following list in order of preference: IDP, SFB, Wittgenstein Source, WiTTFind, Wittgenstein XML TEI, Wittgenstein OWL.
12) How likely is it that you would recommend WAB resources to your students? Please answer with a number between 1 and 10. (10 – extremely likely. 1 – not at all likely)

13) If you recommend WAB resources to your students, which of the following would you recommend? Please choose from the following list in order of preference: IDP, SFB, Wittgenstein Source, WiTTFind, Wittgenstein XML TEI, Wittgenstein OWL.
14) How likely are you to follow WAB on social media such as Facebook, supposing that, amongst other things, this would provide a community platform for discussion of practical and theoretical research questions? (10 – extremely likely. 1 – not at all likely)
15) What is your first and last name?
16) What is your email address?
17) What is your field of research or study? e.g. philosophy, linguistics, digital humanities, text technology, information technology etc.
18) Any other comments/feedback on WAB resources?

4.1. Discussion of survey results

In this section, we will briefly discuss results from the initial phase of the survey, in which we surveyed a small target group of users. This target group consists exclusively of experienced scholars whom we selected on the basis of their long-term participation in the Wittgenstein research community and high degree of familiarity with WAB's digital resources. To date we have received a total of 26 responses. We should stress that we are only reporting a subset of our survey results that we consider to be of special interest, and that we are planning to continue surveying WAB users in this way for the foreseeable future. The survey responses we report in this paper come from participants in a wide range of disciplines, including philosophy, digital humanities, text technology, linguistics, artificial intelligence, philology, literary theory and criticism, learning technologies, information technology, book design, and musicology. The majority of the respondents (21 out of 26) identify their primary research field as being philosophy, often more narrowly qualified as, for instance, philosophy of text, philosophy of artificial intelligence, or philosophy with digital humanities. Overall, the respondents' profiles reveal the wide-reaching impact of digitising Wittgenstein's works.

Answers to question (1), "How did you first hear about WAB resources?", reveal an active interest amongst respondents in searching online for digital Wittgenstein resources. Several respondents said they first came across WAB's digital resources as a result of Google searches. We did not ask these respondents what specific search terms led them to WAB, but this question will be included in future surveys. Other respondents said that they first heard about the WAB resources through presentations at academic events that mentioned WAB or its digital resources.

We will discuss the answers to questions (2) and (3) in more detail.

Question (2) was "Which of the following digital tools are you familiar with or have you used? IDP, SFB, Wittgenstein Source, WiTTFind, Wittgenstein XML TEI, Wittgenstein OWL." Answers to this question indicate that the respondents tend to be especially familiar with IDP and Wittgenstein Source. Twenty-five respondents said that they were familiar with at least one of IDP or Wittgenstein Source. Specifically, 18 said they were familiar with IDP, and 17 that they were familiar with Wittgenstein Source.

Wittgenstein Source provides open access to primary sources. In particular, it includes the *Bergen Nachlass Edition* (BNE), which is the latest digital "combined" static edition of Nachlass facsimiles and transcriptions. IDP, on the other hand, provides a complete digital interactive edition. In section 3 above, we argued that IDP represents an important step forward in the digitisation of Wittgenstein's works. The survey responses to question (2) support this claim, insofar as they indicate the relative popularity of IDP amongst leading experts.

Regarding the semantic tools, 18 respondents said they were familiar with WiTTFind; 16 were familiar with SFB; and nine were familiar with Wittgenstein OWL.

Question (3) was "Which digital tool do you use the most?" Answers to this question indicate that a relatively large number of surveyed users are familiar with the semantic tools but do not actually use them. Twelve respondents (i.e. 46.2%) said they used Wittgenstein Source the most, compared with eight (30.8%) for IDP, four (15.4%) for WiTTFind, one (3.8%) for SFB, and one (3.8%) for Wittgenstein OWL.

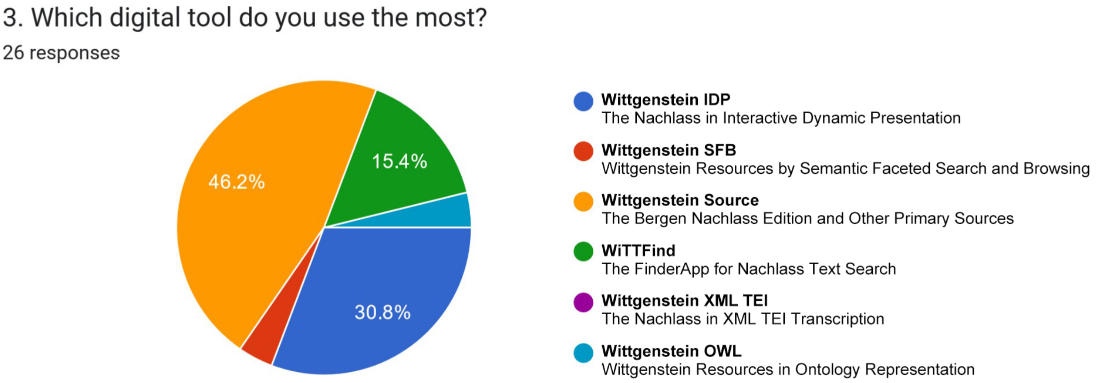


Figure 2: Responses to question (3), "Which digital tool do you use the most?"

This distribution of responses does not come as a surprise to us. Users with purely philological interests may often only require access to a particular facsimile, and this need will be best met by Wittgenstein Source. In addition to Wittgenstein Source being perhaps the most basic resource, it also contains the *Bergen Nachlass Edition* (BNE), an edition that is very similar to the *Bergen Electronic Edition* (BEE). So another plausible explanation for the relative popularity of Wittgenstein Source amongst expert users, as indicated by the reported answers to question (3), is the natural conservative tendencies to keep using the tool that has worked well for you in the past (or at least something that closely resembles it) and not to try using new tools or doing new things. Moreover, in the case of a tool such as the RDF pilot on Wittgenstein OWL, these tendencies may be exacerbated by the amount of technical knowledge that its use currently requires. (Notably, Wittgenstein Source, IDP, SFB, and WiTTFind are reportedly easy to use; see also the responses to question (4) below.)

Of course, users familiar with IDP who wish to work with some specific filtering or sequencing of Nachlass text, for example, can currently only use IDP for this purpose, not Wittgenstein Source or any other digital tool currently available; so users with this kind of need will naturally go and use IDP. Similarly, users familiar with IDP who appreciate its interactive character, by contrast with editions (such as BNE) that offer a static combination of normalised and diplomatic transcription, will naturally tend to use IDP. In general, we expect that research and teaching needs amongst experts will develop in such a way that IDP's relative popularity amongst this group, compared with its static-edition competitor, will continue to grow.

In order to understand these survey results correctly, it is important to note a few more details about the individual histories of the digital tools offered by WAB. For example, users with advanced requirements in terms of combining text strings and semantic search functions, or faceting Nachlass text in relation to metadata, will often find that SFB cannot yet fully satisfy their requirements. The development and dissemination of WAB's various tools followed different timelines and strategies. As it happens, substantial components of many tools became available online in some form or other in quick succession: Wittgenstein Source around 2009, IDP around 2010, WiTTFind around 2011, and SFB around 2012. However, their relative stages of development differed quite significantly. For example, Wittgenstein Source matured relatively early, while SFB took longer to mature than the other tools.

The following statistics may be useful in this context. Google Analytics has recorded over 30,000 Wittgenstein Source users since 2013; over 8,000 IDP users since 2017; and more than 2,000 SFB users since 2017. This comparison confirms the impression that WAB's semantic tools are relatively little used. We believe that this mainly reflects a lack of acquaintance with these kinds of tools amongst users at present, as well as user demand for more sophisticated semantic tools. We have not seen any evidence that would suggest that users are in principle not interested in using digital semantic technologies.

We will conclude our discussion in this section by briefly noting some of the other survey responses.

Question (4) was "Do you easily and usually find what you are looking for when visiting WAB resources?" Twenty-four respondents (92.3%) answered "yes".

Question (5) was "What, if any, are the challenges you face when using WAB resources?" Of the 17 respondents who mentioned some challenge or other, seven indicated that WAB resources were not well presented online, and in particular that it would be better to have a comprehensive landing page with a clear overview of all available resources. WAB has since tried to satisfy this demand by launching its new portal wittgensteinonline.no.

Question (8) was "Briefly describe the advantage(s) of using a digital tool in your research. Did it allow you to ask different research questions? Did it allow you to complete a research process more quickly or efficiently? Did it help in some other way?" Nineteen respondents said that using a digital tool in their research improved efficiency, with the advantages of general searchability being the most commonly cited reason. One respondent wrote: "Digital tools are immensely helpful to identify Wittgenstein's treatment of certain concepts at certain times or periods in his work. By using them one may accomplish research results more quickly and efficiently." Another wrote: "Excellent for searching manuscripts & typescripts and comparing to the edited volumes." Yet another wrote: "Chronological and other sorting as well as filtering of data are only reasonably possible with encoded digital resources; I use IDP and SFB a lot for that."

Question (10) was "How likely is it that you would recommend WAB resources to colleagues?" Of the 26 respondents who answered this question, 22 (84.6%) gave a score of 10 (extremely likely).

Question (12) was "How likely is it that you would recommend WAB resources to your students?" Of the 24 respondents who answered this question, 14 (58.3%) gave a score of 10 (extremely likely). The contrast between the responses to questions (10) and (12) is perhaps not that surprising. Respondents who gave a score below 10 for the question about students also gave plausible reasons for that score, such as wanting to encourage students to focus on key texts first (especially the *Tractatus* or *Philosophical Investigations*) or the fact that many of their students lack proficiency in German. WAB has been exploring various options for integrating translations into its digital tools, and it is expected that more resources will be made available in English in the future. In future surveys, we plan to collect more data about the use of WAB's digital resources by students, and we plan to include additional questions about teaching-related user needs.

5. Conclusion

In previous sections, we discussed the nature and history of, and future directions for, the digitisation of one of the most complex philosophical oeuvres of modern times. In particular, we discussed how and why the digitisation process conducted at the Wittgenstein Archives at the University of Bergen (WAB) has been informed, and in many ways fundamentally driven, by the complex needs of its user community. The growing complexity of user needs is largely a function of the complexity of the texts and a growing awareness of the potential of digital humanities. Wittgenstein's philosophical oeuvre, much like the man himself, resists being fitted into any standard intellectual or technological mould. It therefore presents some unique opportunities and challenges, which may be used to explore new directions in digital humanities.

WAB's development of digital resources over the past three decades illustrates various ways in which digital resources can be superior tools to traditional paper editions for research in the humanities. WAB has thus helped to establish a large and growing global community of students and scholars of Wittgenstein's philosophy. Today, this community's needs for digital tools and resources are becoming increasingly complex. These evolving needs must guide WAB's ongoing and future development of digital tools and resources for the study of Wittgenstein's philosophy.

We would like to end by acknowledging some of the encouraging responses to question (18) of our survey: "Any other comments/feedback on WAB resources?" Responses included "Thank you for [the] great service", "The WAB resources are wonderful in every way", "Where would we be without them?", "Love the archives online", and "Keep up the great work!"

Acknowledgments

The authors would like to thank Frank Liang for his assistance with preparing the final version for publication.

References

Christian Erbacher. *Wittgenstein's Heirs and Editors.* Cambridge University Press, 2020.

Max Hadersbeck, Sabine Ullrich, Ines Röhrer, Sebastian Still, and Alois Pichler. Spielräume bei der retroperspektivischen Analyse der Wittgenstein-Edition und die Herausforderungen für das Semantic Clustering. In *7. Jahrestagung der Digital Humanities im deutschsprachigen Raum (DHd 2020).* Paderborn, 2020.

Herbert Hrachovec. Wittgenstein on line / on the line, 2000. URL https://wab.uib.no/wab_contrib-hh.page.

Claus Huitfeldt. Toward a machine-readable version of Wittgenstein's Nachlaß. In Hans Gerhard Senger, editor, *Philosophische Editionen: Erwartungen an sie – Wirkungen durch sie. Beiträge zur VI. Internationalen Fachtagung der Arbeitsgemeinschaft philosophischer Editionen*, pages 37–43. De Gruyter, 1994.

Federico Meschini. Oltre il libro: Forme di testualità e digital humanities. *Oltre Il Libro*, pages 1–296, 2020.

Danièle Moyal-Sharrock, editor. *The third Wittgenstein: The Post-Investigations Works.* Ashgate, 2004.

Alois Pichler. Ludwig Wittgenstein and us 'typical Western scientists'. In Sebastian Sunday Grève and Jakub Mácha, editors, *Wittgenstein and the Creativity of Language*, pages 55–75. Palgrave Macmillan, 2016.

Alois Pichler. Complementing static scholarly editions with dynamic research platforms: Interactive Dynamic Presentation (IDP) and Semantic Faceted Search and Browsing (SFB) for the Wittgenstein Nachlass. In *CLARIN Annual Conference*, pages 194–207, 2021.

Alois Pichler and Tone Merete Bruvik. Digital critical editing: Separating encoding from presentation. In Daniel Apollon, Claire Bélisle, and Philippe Régnier, editors, *Digital Critical Editions*, pages 179–199. University of Illinois Press, 2014.

Alois Pichler and Odd Einar Haugen. Fra kombinerte utgaver til dynamisk utgivelse: Erfaringer fra edisjonsfilologisk arbeid med Wittgensteins filosofiske skrifter og nordiske middelaldertekster. In *Læsemåder: Udgavetyper og målgrupper*, Nordisk Netværk for Editionsfilologer, Skrifter, pages 178–249. Reitzels Forlag, 2005.

Alois Pichler and Amélie Zöllner-Weber. Sharing and debating Wittgenstein by using an ontology. *Literary and Linguistic Computing*, 28(4):700–707, 2013.

Alois Pichler, James M. Fielding, Nivedita Gangopadhyay, and Andreas L. Opdahl. Crisscross ontology: Mapping concept dynamics, competing argument and multiperspectival knowledge in philosophy. In Fabio Ciracì, Richard Fedriga, and Cristina Marras, editors, *Quaderni di "Filosofia", Number 2: "Filosofia digitale"*, pages 59–73. Mimesis, 2021.

Ines Röhrer. Lexikon, Syntax und Semantik – Computerlinguistische Untersuchungen zum Nachlass Ludwig Wittgensteins. Master's thesis, LMU Munich, 2019.

David Stern. The 'middle Wittgenstein': From logical atomism to practical holism. *Synthese*, 87 (2):203–226, 1991.

David Stern. How many Wittgensteins? In Alois Pichler and Simo Säätelä, editors, *Wittgenstein: The Philosopher and his Works*, pages 205–229. Ontos Verlag, 2006.

Sebastian Sunday Grève. The importance of understanding each other in philosophy. *Philosophy*, 90(2):213–239, 2015.

Sebastian Sunday Grève. Logic and philosophy of logic in Wittgenstein. *Australasian Journal of Philosophy*, 96(1):168–182, 2018.

Sebastian Sunday Grève. Real names. In Martin Gustafsson, Oskari Kuusela, and Jakub Mácha, editors, *Kripke and Wittgenstein.* Routledge, forthcoming.

Alfred North Whitehead and Bertrand Russell. *Principia Mathematica.* Cambridge University

Press, 1910.

WiTTFind: A cooperation between the Wittgenstein Archives at the University of Bergen under the direction of Alois Pichler and the Center for Information and Language Processing at LMU Munich under the direction of Max Hadersbeck [wittgensteinonline.no]. Bergen and Munich 2016–.

Ludwig Wittgenstein. *Tractatus Logico-Philosophicus.* Kegan Paul, 1922.

Ludwig Wittgenstein. *Philosophical Investigations*, edited by P. M. S. Hacker and Joachim Schulte. Wiley-Blackwell, 4th revised edition, 1953/2009.

Ludwig Wittgenstein. *The Blue and Brown Books: Preliminary Studies for the 'Philosophical Investigations'*, edited by Rush Rhees. Harper & Row, 2nd edition, 1958/1969.

Ludwig Wittgenstein. *On Certainty*, edited by G. E. M. Anscombe and G. H. von Wright. Blackwell, 1969.

Ludwig Wittgenstein. *Philosophical Occasions: 1912–1951*, edited by James C. Klagge and Alfred Nordmann. Hackett, 1993.

Ludwig Wittgenstein. *Wittgenstein's Nachlass: The Bergen Electronic Edition.* Oxford University Press, 2000. [BEE]

Ludwig Wittgenstein. *Wittgenstein Source Bergen Nachlass Edition*, edited by the Wittgenstein Archives at the University of Bergen under the direction of Alois Pichler. In *Wittgenstein Source* (2009–) [wittgensteinonline.no]. Bergen, 2015–. [BNE]

Ludwig Wittgenstein. *Interactive Dynamic Presentation of Ludwig Wittgenstein's Philosophical Nachlass*, edited by the Wittgenstein Archives at the University of Bergen under the direction of Alois Pichler [wittgensteinonline.no]. Bergen, 2016–. [IDP]

Wittgenstein OWL: The Wittgenstein Domain in Ontology Representation, edited by the Wittgenstein Archives at the University of Bergen under the direction of Alois Pichler [wittgensteinonline.no]. Bergen, 2006–. [OWL]

Wittgenstein SFB: Wittgenstein Resources by Semantic Faceted Search and Browsing, edited by the Wittgenstein Archives at the University of Bergen under the direction of Alois Pichler [wittgensteinonline.no]. Bergen, 2012–. [SFB]

Alan Wood. *Bertrand Russell: The Passionate Sceptic.* Allen and Unwin, 1957.

Georg Henrik von Wright. *Wittgenstein.* Blackwell, 1982.

(R)Unicode: Encoding and Sustainability Issues in Runology

Elisabeth Maria Magin[1], Marcus Smith[2]

[1]*Museum of Cultural History, University of Oslo*
[2]*Swedish National Heritage Board*

Abstract
In this article, the basic premises of reading a runic inscription and transferring the data into digital formats are discussed. Particular attention will be on the Runic block in the Unicode standard for digital character encoding, which currently does not suit the needs of the academic runological community. The proposed solution builds upon the existing standard without adding unnecessary additional characters, but with the ability to encode form-variants on top of the base character while retaining backwards compatibility, an approach from which other archaic scripts like Cuneiform or Khitan that suffer of similar encoding issues, could also benefit.

Keywords
Runes, Unicode Runic, Runology, Ancient scripts, Digitisation

1. Introduction

While most commonly associated with the Vikings in the popular consciousness, the term "runes" is used for at least four closely related variations of the same alphabetic writing system used across a geographic area from Ukraine to Greenland, from the 3rd to 19th centuries CE. The lion's share of surviving runic inscriptions are found in Sweden, most often carved into stones, and Norway, where more than half of the extant runic corpus was carved into small everyday objects, often wood or bone. During the approximately 1700 years of active use, the original 24-character runic row referred to as the Older Fuþark (alternatively futhark) underwent several major changes. In the British Isles, the character inventory was expanded (Anglo-Saxon Fuþorc), while in Scandinavia, only 16 runes of the original 24 remained in use during the Viking Age (Younger/Viking-Age Fuþark). During the High Middle Ages, these 16 runes were modified and the inventory expanded to properly represent the phonemes in later Old West Norse (Younger medieval Fuþork). There is also a post-Reformation runic tradition, which is, however, not the topic of this article [introductions in 1, 2, 3].

Studying the main variations of runic writing and their development and derivatives remains however challenging even with digital methods to support research, a problem also frequently

DHNB2023 | Sustainability: Environment - Community - Data. The 7th Digital Humanities in the Nordic and Baltic Countries Conference. Oslo – Stavanger – Bergen, Norway. March 8–10, 2023.

e.m.magin@khm.uio.no (E. M. Magin); marcus.smith@raa.se (M. Smith)
https://www.khm.uio.no/english/about/organisation/archaeology-department/staff/elismari/index.html (E. M. Magin)
0000-0002-7612-1234 (E. M. Magin); 0000-0002-7438-5498 (M. Smith)

DHNB Publications, DHNB2023 Conference Proceedings, https://journals.uio.no/dhnbpub/issue/view/875

encountered by scholars working with other archaic scripts like Cuneiform or Khitan. In recent years, this has become a serious drawback in runology, as digital searches would for example enable runic scholars to conduct macro- instead of the currently most common microstudies of runic inscriptions. It is currently also impossible to search runic inscriptions *as texts* without relying on the current workaround, transliterations into Roman letters.

Getting the runes into the computer has, however, never been a straightforward endeavour. Several of the difficulties encountered can be attributed to encoding runic characters using the Runic block, added to the Unicode standard in 1999 to allow digital representation of runes. The initiative was driven by several established and senior runologists, for example Helmer Gustavsson (Sweden), James Knirk (Norway), Klaus Düwel (Germany), Ray Page (Great Britain) and Marie Stoklund (Denmark) [4]. It was only accepted as a revised proposal, which led to a code block Runic being created, then featuring 81 runes from the Older Germanic, the Anglo-Frisian, the Viking Age and the medieval Fuþarks. It also includes characters of what are generally considered derivatives of runic writing, referred to as "rune rows", in this case the long-branch and short-twig variations. These appear as allographs during the Viking Age, although some of them are later distinguished and become separate characters in the medieval Fuþark [for example 5, 6].

As this short introduction exemplifies, the development and reciprocal influence of different runic scripts is by no means clean-cut and unambiguous. Within the field, there is general agreement that the standardised Fuþarks presented in handbooks and introductions to runes are simplifications of a far more complex situation [7]. They certainly serve a purpose [7, 6-8] and one may argue that character encoding and Unicode is an area where simplification and standardisation are precisely what is required. The original code block Runic (not the later additions from the Franks Casket and Tolkien runes, which have caused much consternation to those working with runes professionally – in the first case, the runes are allographs of already encoded characters, in the second, they are part of a fantasy script, not actually part of any historical runic script) to a certain extent follows these principles. The code block, due to inconsistent encoding, still leaves much to be desired for the runologist, however, especially when the aim is to conduct research into the genesis and use of the different runic rows; as such its use is mostly confined to non-runologists. It is the authors' contention that the Unicode Runic block does not serve the needs of runic scholars, following no clear principles in which characters are encoded and which are considered as allographs and therefore not encoded.

Following a short introduction into the differences between the four main runic rows on a graphemic and linguistic level, this article proceeds to the different stages in the process of deciphering a runic inscription. It will explain how runologists currently remedy the issues with Unicode Runic when encoding runic inscriptions and why these are unsustainable, before examining a potential solution making use of Unicode Form Variation Selector and Stylistic Sets in OTF fonts in an attempt at a more sustainable approach to runic encoding better aligned with the needs of runic scholarship in the twenty-first century. Since most runologists work with inscriptions belonging to one of these four Fuþarks, focus here rests on the encoding of these, although runes were actively used in some areas until the 18th century.

2. Fuþarks and rune rows: runic writing systems

The first potential, albeit debated, runic inscription is a fibula from Meldorf, Germany, dated to the first half of the 1st century AD. Equally debated other early inscriptions originate from Vimose, Denmark, and date to the 2nd and 3rd centuries AD. The currently most widely accepted explanation is that the Older Fuþark was invented somewhere in the regions of Northern Germany or Denmark at some point during the first two centuries AD, inspired by the Roman or other Italic alphabets, and then used in areas settled by Germanic tribes. This oldest version consists of 24 signs divided into three ǽttir with eight runes each: ᚠᚢᚦᚨᚱᚲᚷᚹ•ᚺᚾᛁᛃᛇᛈᛉᛊ•ᛏᛒᛖᛗᛚᛜᛞᛟ

These stayed in use until around 800 AD, when the character inventory was (perhaps gradually) reduced to just 16 runes around beginning of the Viking Age on the continent [e.g. 8]. In Britain, to where the Angles, Saxons and Jutes had exported the Older Fuþark, the opposite development took place: the inventory was gradually expanded to 32 runes owing to changes in the spoken language, which gained more distinct speech sounds. No scholar has been able to find an explanation for these contradictory developments, since the language changes on the continent also led to an increase in distinct speech sounds [e.g. 3]. As runes are a phonemic script, meaning each character (grapheme) is tied to a distinct speech sound or sounds in spoken language (phoneme), the reduction of the rune inventory on the continent during the Viking Age continues to be a topic of scholarly debate. This situation is compounded by the development of no less than three derivatives of the Younger Fuþark, referred to as long-branch, short-twig and staveless runes. None of these are considered to be separate Fuþarks, instead the runes are considered to be allographs in derivative systems, the reasons for which become obvious when comparing them:

Long-branch: ᚠᚢᚦᚬᚱᚴ : ᚼᚾᛁᛅᛋ : ᛏᛒᛘᛚᛦ   Short-twig: ᚠᚢᚦᚭᚱᚴ : ᚽᚿᛁᛆᛌ : ᛐᛓᛙᛚᛧ

This approach is also owed to the fact that runecarvers could freely choose which variation they were using; inscriptions frequently mix and match the systems, excepting staveless runes [9, 10]. Why runecarvers would make use of characters from either system or whether they even considered them to be distinct systems is not fully understood.

Disregarding developments of runic script post-15th century, one more change in the runic writing system takes place around c. 1100 with the end of the Viking Age. Potentially owing to the problem of having to use the same rune to express several different speech sounds and the resulting confusion, the Medieval Fuþark starts to de-unify runes and adopts, for example, long-branch ᛅ and short-twig ᛆ as separate characters. New runes are also added to the repertoire [for example 11]. Use of these extra characters is, however, not mandatory; the characters in an inscription archaeologically dated to the middle of the 13th century AD can therefore look precisely the same as those in an inscription dating to the 9th century AD.

3. The three (or more) stages of reading a runic inscription

As described in a variety of (runic) handbooks [12, 1, 2, 3], the reading or interpretation of a runic inscription generally proceeds in three steps following an initial examination. The first of these – unless drawings and images are used instead – is most often referred to as “transcription” or “transrunification” and describes the process of standardising the shape of

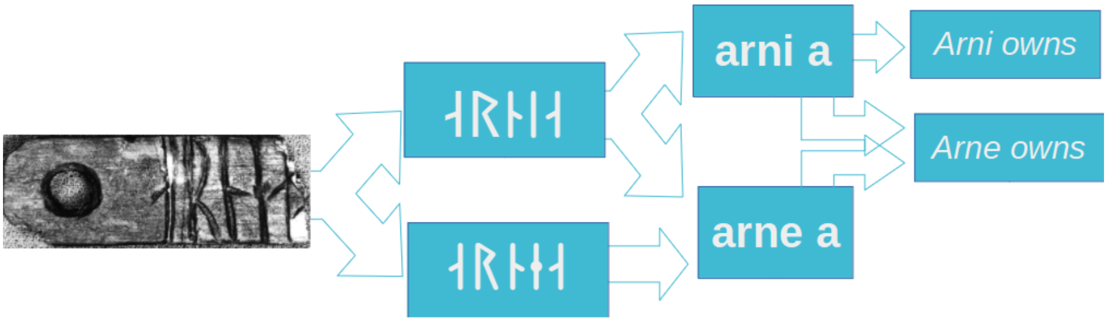


Figure 1: The three stages of interpreting or reading a runic inscription

the runes on an object (more rarely in a manuscript) into what various authors have referred to as "idealruner" [9] or "print runes" as illustrated by the second "column" in 1. At its most basic, this is the same process as typing up handwritten notes on a keyboard, just with runes instead of Roman letters. However, other than the well-standardised Roman alphabet and despite the fact that standardised Fuþarks exist, transcribing runes is somewhat more complicated since runic writing exhibits a variety of peculiarities that frequently render the process less than straightforward. Firstly, while the direction of writing is generally left-to-right, inscriptions in the Older Fuþark in particular can be written left-to-right or right-to-left, or sometimes both in alternating lines, a phenomenon called "boustrophedon" (as the ox plows).

To complicate matters further, even single runes within one inscription can be carved against the prevalent writing direction, becoming "Wenderunen" (turned/flipped/reversed/inverted runes). This happens as late as the Middle Ages, where a prominent example is N 737 [13], reading ᛐᛁᚢᛐᚱ, interpreted either as the name Ljótr or Þjóðarr. If the latter is correct, the use of ᛐ instead of ᚦ is a little surprising, but not impossible. However, ᛐ, the **t**-rune, can also be interpreted as ᛚ, the **l**-rune, written against the otherwise prevailing left-to-right writing direction in the inscription. The same phenomenon can be observed for ᚿ (short-twig **n**) and ᛆ (short-twig **a**), ᚾ (long-branch **n**) and ᛅ (long-branch **a**), ᚬ (**o**) and ᚮ (**o**) and others, for example ᚴ (**k**), which can be flipped upside-down: ᛧ. The impact of this phenomenon on reading and interpretation varies. As N 737 shows, a Wenderune can lead to a completely different reading of a character sequence, whether that be a name or another type of word. Other Wenderunen can appear both as Wenderunen and as characters in their own right: ᚬ (**o**), when turned around, is also used to spell the letter "b" in Viking Age and medieval inscriptions. At this point, runologists are already faced with a difficult decision: in print, do they represent the runes as they *see* them or as they *interpret* them?

Seim [1] is not the only runologist arguing for a strict division between describing the purely visual and the interpretation of what the scholar sees. A variety of publications, for example Waldispühl [14], Palumbo [6], Nowak [15], Spurkland [16] analyse the relationship between graphemes, graphtypes, graphtype variants and glyphs. The studies, for the most part, define "grapheme" as the abstract character a given sign is supposed to represent, which is in turn connected to a specific sound value in phonemic scripts like runes. Graphtypes represent the idealised visual representation of that abstract character, glyphs are the actual characters on objects or in manuscripts, basically the “handwritten” forms that are subject to a range of influences. Graphtype variants group the “handwritten” runes together based on visual similarities, roughly equivalent to allographs; they represent a level between the single actual written runes and the idealised graphtypes. For example, during the Viking Age, ᛅ and ᛆ could be considered allographs or graphtype variants of a graphtype ᛅ, with ᚾ and ᚿ being further Wenderunen allographs of this graphtype, while in themselves also representing a separate graphtype ᚾ with the short-twig allograph ᚿ. This shall be returned to in 4.

The second step is usually called “transliteration” and describes the process of equating runes with Roman letters. Again, practices can differ. Some runologists equate rune=Roman letter by way of the phoneme or sound value of the respective rune/letter, whereas others use Roman letters as a stand-in for the graphical form of the rune which may also signify a letter-sound relationship, but does not have to [17, 307]. It is illustrated in the second column in 1. The main problem at this stage arises from the fact that the speech sounds Roman letters are ascribed and the reconstructed speech sounds for runes do not always align – and that in specific Fuþarks, one and the same rune can be used to express several different speech sounds. One example of this is illustrated in 1, where the first sequence of runes can be transliterated as both **arnia** and **arnea**, since ᛁ is a multi-valued rune used primarily to express the speech sound /i/, but with a secondary value /e/. However, around c. 1100, several multi-valued runes acquire a diacritic marker, a dot, to differentiate between the primary and the secondary sound value. This is illustrated in the first column, where the first transcription shows ᛁ while the second uses ᛂ (since the dot is usually created by simply sticking the point of whatever sharp object – mostly a knife – is used to carve the inscription, it can be fairly hard to decide whether a rune is dotted or not). This rune is unambiguous inasmuch as it is never used for /i/, so when it appears in an inscription, interpretation is fairly clear. Technically, this would solve some problems for modern rune-readers and runologists *if* the system of using dots as diacritic markers to distinguish between primary and secondary sound value was used consistently – which it was not. Instead, using the system was up to the single runecarver’s preferences.

Another complicating factor arises in the fact that the short-twig **h**-rune looks exactly alike. So interpretation is clear only when ᛂ appears in a context where interpretation as short-twig ᚽ would make no sense, but reading it as /e/ does.

As mentioned above, the phenomenon of multi-valued runes is particularly observed in the 16-character Viking Age Fuþark, although since use of the system is optional, the problem continues in medieval inscriptions. The diacritic dot also fails to solve the problem when a rune has more than a primary and secondary sound value, like ᚢ. In this case, the dot generally only serves to exclude the primary sound value from consideration, in this case /u/. ᚤ can still signify /o/, /y/, /v/, /w/ or /f/.

Different runologist communities follow different traditions regarding transliterations to handle the problem. Today, many runologists follow the custom of transliterating runes with the same Roman letter regardless of supposed sound value in any given instance [for example 18, 105]. This, however, was not the case for earlier publications or, for example, many German runologists, whose transliterations frequently aim to represent the suspected pronunciation of any given inscription rather than following a consistent transliteration system. This leads to the interesting phenomenon that a transliteration of ᛁᚢᛆᚱ may look like **iuar** when transliterated by a runologist following the first approach, but **ioar** when transliterated by one following the latter. Furthermore, the sequence may then be normalised by the first as *Ívarr*, while the second may instead choose to normalise as *Jóarr* – which are two different names with different etymology and meaning (although they can also be variants of the same name) [19, on different transliteration conventions see].

This last step of the process, normalisation, mainly aims to render the text of the runic inscription according to the modern-day standards of spelling the different languages appearing in runic inscriptions, for example proto-Norse, Old East Norse or Old West Norse. The main issue here arises once more precisely from the fact that multi-valued runes can be interpreted as a variety of speech sounds. This in turn can lead to differences in final conclusion like Ívarr/Jóarr or, as in 1, a reading of the name as Arni versus Arne. While this may in the first instance seem to be a minor issue, for Arni and Arne are, after all, just two spellings of the same name, this uncertainty can become a much greater problem when, for example, linguists attempt to use runic inscriptions to trace and date sound changes in a given language, as is the case for the evolution of Old West Norse into Middle Norwegian between 1350 and 1550. Although scholars mainly use manuscripts for the purpose, indications for the change also appear in runic inscriptions, specifically the inscriptions found in Bryggen, Bergen, Norway [13]. Encoding runic inscriptions in such a way that linguists can use the resulting data as well, while retaining the information necessary for them to decide whether a particular inscription is relevant to their study, is therefore not just a desideratum, but a requirement.

4. Current encoding solutions

This, however, is where the problems begin, since runologists are more often than not interested not only in runes as strings of characters, but in what the runes in any given inscription look like. It should be clear from 2 that the specific form of a rune carries importance, whether in terms of dating, the potential origin of the runecarver or simply just to decide which particular speech sound the rune represents. Runologists are therefore not only in need of *one* transcription; in theory (although this is currently not established as a custom), they are in need of transcriptions at *two* levels, one simply representing the purely visual, the second representing the interpretation of the purely visual. The code block Runic in the Unicode character encoding, despite being developed by runologists, is unfortunately inconsistent in how and which runes are encoded, more often than not mixing the different levels, which has led to the code block not being used at all by the academic runologist community.

Instead, they currently fall back on one of two solutions to remedy the issues with Unicode Runic. The first is to ignore the runes as characters completely, instead presenting only a transliteration into Roman characters, supplemented with photographs or drawings of the inscription. This approach is complicated not only by the different approaches taken to transliterations within the academic runologist community, but also the fact that this means the runes are often not represented as runes *at all* in the given publication. For someone specifically aiming to examine the use of different graphtypes and graphtype variants, this approach is therefore less than useful.

Alternatively, runologists rely on bespoke typefaces allowing them to represent a broader range of the visual forms of runes more accurately. This does allow for representation of both graphtypes and graphtype variants, but creates another problem, namely at which level of accuracy the runes are represented (3). Runic scholars rarely specify whether they are using a purely visual approach and represent the runes in print exactly as they see them on the object, Wenderunen and graphtype variants included, or whether they are choosing to print their interpretation of what it is they see, using graphtypes instead of graphtype variants, or even a mix of both approaches. It is generally through the comparison with the images provided that one is able to determine which kind of transcription one is dealing with.

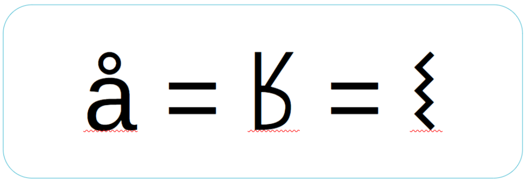


Figure 2: U+00E5 printed using a Unicode font and the non-Unicode runic fonts Gullskoen and Gullhornet

Neither method is particularly sustainable in the long run. Images for many runic inscriptions are often lacking completely or of poor quality and hindered by copyright issues, and transliteration customs vary from country to country and by runic writing system. Custom runic typefaces rely on everyone making use of the data having access to the same typeface, as the underlying characters remain encoded as plain ASCII, unreadable without the font. This is illustrated in 2, where the image of the character å, which should be encoded by the codepoint 00E5, has simply been replaced by, in the Gullskoen font, a flipped version of the **r**-rune, whereas the Gullhornet font uses the same codepoint for a variation of the Older Fuþark **s**-rune.

While acceptable for print purposes, this solution is ill-suited for long-term storage purposes, the documents thus created are virtually unsearchable since they encode runic characters as Roman letters. Even worse, if more than one font was used for document and the fonts are inconsistent in their use of codepoints, like Gullskoen and Gullhornet, anyone trying to conduct a document search has to basically know which fonts were used for document creation and which codepoints they respectively use for different runic characters.

But even if one were to accept the use of already existing, bespoke codepoints for characters from a completely different script, the use of fonts and graphtype variants creates other issues for the kind of analyses runologists or linguists may want to conduct. While the visual form of a rune is highly relevant to especially graphemic analyses of runic corpora, for analyses of combinations of different runes or potential changes in the spoken language, the precise visual form is irrelevant to begin with. However, if fonts are used at the graphtype variant level, this means that every graphtype variant is also encoded using a separate codepoint. In turn, this means that to find and compare inscriptions including the same word or name, all different possible combinations of the graphtype variants need to be searched for, a task that quickly becomes unmanageable when single graphtypes can span 10 or more graphtype variants. Perhaps this would not be as much of a problem as it is if the different runic scholarship traditions did not follow different customs where transliteration and normalisation are concerned. The situation is what it is, however, and from a purely methodological perspective, it is also inadvisable to conduct searches on what are higher-level interpretations of a piece of writing, which transliterations and normalisations are.

The question may be asked why runologists do not make use of Unicode Runic instead of fonts then. The simple answer to that is that Unicode Runic currently suffers of precisely the same problem as fonts in that respect that it is inconsistent in how graphtypes and graphtype variants are encoded. For example, while there is an argument to be made that ᛅ and ᛆ do represent different graphemes starting around c. 1100 AD, when ᛅ starts to be used to represent /æ/ and ᛆ continues to represent /a/, long-branch ᚾ and short-twig ᚿ are also each given their own codepoint, completely disregarding the fact that either rune always and consistently (unless it is a Wenderune) represents /n/.

Other issues arise from the fact that some of the characters come with an inherent dating and geographical placement. Older Fuþark ᚨ and Anglo-Frisian ᚫ look exactly alike, the only argument for encoding them as separate codepoints being that the latter also comes to represent /æ/ from a certain point in time. When precisely is still a topic of discussion amongst runologists; but even if that were not the case, objects and people were mobile. Using the Anglo-Frisian ᚫ to encode an inscription that could also have been carved by someone using the Older Fuþark – far from an unlikely scenario, the two Fuþarks being in parallel use and the geographic areas being in contact – automatically hardcodes an interpretation in what *should* be a neutral description of a visual sign. Encoding at the graphtype level by using Unicode Runic is therefore problematic from a methodological point of view; for encoding at the graphtype variant level, however, Unicode Runic simply lacks too many runes, and this approach additionally results in the same problem as using fonts does by giving each graphtype variant its own codepoint.

To the best knowledge of the authors of this article, the academic runologist community have therefore avoided actively using Unicode Runic whenever possible. There are either too many or too few encoded characters to be of practical use for the kinds of analyses runologists and scholars from adjacent fields like linguistics or archaeology would like to run.

It is instructive here to compare the digital encoding of runes with similar ancient and epigraphic scripts. Perhaps the closest ancient script is the insular Celtic Ogham, in use from around the 4th century CE, and continuing in scholastic use into the 9th century. Like runes, Ogham forms were optimised for carving on wood, metal, stone, or bone, with orthogonal and diagonal tally-like forms. The simple and uniform nature of Ogham does not lend itself to orthographic variation, and allographic deviations are rare. The Ogham Range was added to Unicode in 1999, in the same revision as Runic, and provides adequate coverage for the digital encoding of Ogham texts. Examples of current digital Ogham corpora successfully employing Unicode include *OG(H)AM* from the University of Glasgow, and the Research Squirrel Engineers' linked open data research hub [20]. In other cases, however, the present state of digital script encoding is in a similarly unsatisfactory state to that of runes, and potential solutions may be applicable to both. This applies to Cuneiform, which (much like runes) is an umbrella term for several variants of the same writing system used to write different languages and showing temporal and geographical variations, and Khitan [21, providing an overview of the encoding process].

4.1. Problems with Unicode Runic

In conclusion of the above, Unicode Runic range and its character coverage are – on their own, and in their current form – not fit for the purpose of transcribing runic inscriptions. In addition to inconsistencies making the correct method of accurate and consistent text encoding uncertain or in some cases impossible, there are also a number of aspects of the standard that are to be found wanting. Firstly, there are a number of runes that are not covered by the standard, which it is thus not possible to reliably encode at the moment. Examples include the runes of the post-reformation Dalecarlian tradition (*Dalrunorna*) used in inscriptions in central Sweden (and, in one notable case, in North America) up until the late 19th century, as well as important variants from the Viking-Age and medieval rune-rows such as the so-called "Greenlandic"-**r** (which in fact occurs just as frequently in non-Greenlandic inscriptions from the period). Also awkward to encode are the staveless runes, a form of the Younger Fuþark frequently used as a form of runic short-hand on ephemeral inscriptions on wood and bone, as well as some monumental inscriptions, especially in the Hälsingland province of Sweden. In the case of the staveless runes, they have in fact been deliberately omitted from the standard, and are considered instead – correctly, in the authors' opinion – to be allographs of the runes of the Younger Fuþark, and should thus be encoded using the short-twig characters of that rune-row, and displayed using an appropriate font [22, ch. 13 Archaic Scripts, specifically 13.3 Runic, 342]. However, this intention is not clear from the code block itself nor widely known among runic scholars: at least one proposal to encode the staveless runes separately has been submitted to the Unicode Consortium apparently in ignorance of this, and rejected [23]. A related problem is that it is not always clear to working runologists not intimately acquainted with Unicode which character should be used in the first place even for characters that do exist. The glyphs provided on the code charts are representative examples only, but for a runologist wishing to encode a short-twig **o** from an inscription in the Man-Jæren group, it would be easy to mistakenly encode a **b** instead, since the forms of those runes are swapped from their usual values in many of those inscriptions.

In addition to runic characters absent from the standard – thus unencodable – which characters *are* encoded, and how, is frequently inconsistent. The long-branch and short-twig allographs of the Younger Fuþark are encoded as separate characters, but as noted above, the staveless allographs are not. It would have been preferable to have treated all such allographs in the same way, either encoding them all or, ideally, encoding only a single character for each and treating the allographs as variant forms. In some cases, separate encodings do make sense, such as the above-mentioned long-branch ᛅ and short-twig ᛆ (4); but in most cases there is no such justification for allographs being allocated distinct codepoints, such as single- and double-barred varieties of Older Fuþark **h** (ᚺᚻ). For other runes, significant allographs are not given their own code points: the closed "maskros" form of Younger Fuþark ᛉ is absent, as is the aforementioned "Greenlandic"-**r**. Once again, some sort of consistent approach would have been preferable here: either encode such variations for all runes, or for none. Perhaps most egregiously, some variant forms occurring in a single inscription have been assigned their own codepoints. This is the case for the five variants of the Anglo-Saxon **o**, **i**, **e**, **a**, and **æ** runes which appear on the Franks Casket [24]. Other such idiomatic glyph variants are not afforded the same luxury; thus, for example, it is not possible to encode the elaborately embellished forms which occur on IMM MM111 Andreas V and GR 1 Kingittorsuaq. Other aspects of encoding the runic script not directly related to character coverage also remain problematic and are largely unaddressed by the Unicode standard; examples include bind-runes (runic ligatures, which might be better modelled a Zero-width Joiners character sequences), and mirrored, flipped, or inverted runes.

4.2. Transcription, glyphs, graphemes (and phonemes?)

As explained above, the grapheme is almost always important, and indeed it is some approximation of the grapheme, guided by a supposed sound value, that is recorded in runic transliteration. Furthermore, often – but not always – the specific glyph that has been used, the form variant, is also important to record. However, as we have seen, in Unicode Runic this distinction between grapheme and glyph inconsistently implemented, making accurate and consistent transcription difficult, sometimes impossible. The rune encoding problem is not really a new one, but rather derives, and is inseparable from, the general problem of transcription and transliteration. On some level, an agreed-upon set of standardised forms (glyphs and glyph-variants) is required, and ideally a set of characters (for our purposes, more or less synonymous with graphemes) that those forms represent. This is complicated by the fact that the correlation between glyphs and graphemes is often not straightforward. Multiple rune forms may denote the same grapheme within a given graphemological system (these are allographs). Conversely, multiple graphemes may be denoted by the same rune forms across graphemological systems (these are homographs). Such distinctions are frequently elided in Unicode Runic. The most accurate representation of a runic inscription – and particularly the forms of the runes themselves – is often a photograph or a drawing, but these are of limited value when trying to encode the texts *as text*. After a graphemic representation, a text transcription is the next level of useful recording with minimal interpretation, and it is this level of recording that Unicode Runic simultaneously seems to seek to enable and fails, disappointingly, to achieve.

So, there is a need in runic transcription to encode not only the grapheme, but also the glyph or form variant. Furthermore, when choosing which runes and rune-forms to encode, there is a choice between taking a maximalist or minimalist approach. With a maximalist approach, one would attempt to assign each identified form variant its own character. With a minimalist approach, one would assign characters only to distinct graphemes, viewing form variants as allographs with distinctions in form to be handled only at the font rendering level. The Unicode Runic range tries to be maximalist in some cases, but minimalist in others, apparently unsystematically, with unsatisfactory results. There are, in any case, issues with both the maximalist and minimalist approaches. While a maximalist character set would allow all identified form variants to be unambiguously encoded at the character level, it would make such encoding, and the use of such an encoded text, tedious indeed. Consider a hypothetical runic database or corpus encoded with Unicode Runic; how would one search for a character sequence in Younger Fuþark inscriptions including the rune **b**? In this case, we're not interested in the form variant of the rune, just the occurrence of the grapheme itself. As things stand at the time of writing, one would have to provide at least two alternatives in the search parameters, to cover cases of either long-branch and short-twig **b** occurring, since these have separate codepoints. But that's just for one grapheme: if we wanted to search for longer strings, the number of necessary alternatives would quickly become unmanageable, as the number of character variations for each rune multiply. In a hypothetical truly maximalist approach, we might have to deal with ten possible alternative forms for **b** alone! Conversely, for the encoder, they would always have to explicitly encode not just e.g. a **b**-rune, but which specific form that **b** took. Being able to encode characters alone, and not the specific graphical representation of each one, is something we take for granted in many other scripts. But there are problems, too, with a minimalist character set, chief among them that it becomes impossible to include information about specific form variants in plain text where that *is* of interest – additional levels of encoding, such as markup or rich text formatting are required on top of the text. So the problem remains: how can we digitally encode both graphemes and form variants for runes and runic texts, without creating separate characters for everything?

5. What can be done?

Having established that Unicode Runic does not live up to the needs of the runologist user community, and given the limitations of the existing Runic range, what might be done to address these shortcomings? There are a few options open to runologists feeling frustrated with this unsatisfactory state of affairs and considering effecting positive change:

5.1. Option 1: Do nothing.

- Accept that Unicode Runic is not suitable for encoding old inscriptions, and should only be used to formulate new texts.
- Rely solely on photos, drawings, transliteration, and some idiosyncratic rune fonts.

This option is not a particularly appealing one, but it is nonetheless a strong possibility, given the limited action runologists as an academic community have taken regarding the standardisation of the script to which they devote their study since the Runic block was first published in 1999. The position that Runic should only be used to formulate new texts is also, sadly, in line with Unicode's original stated remit, explicitly prioritising characters for modern and future use over "preserving past antiquities": "Beyond… modern-use characters, all others may be defined to be obsolete or rare; these are better candidates for private-use registration than for congesting the public list of generally-useful Unicodes" [25, 5]. (One might however speculate that this attitude has softened over the years, given that the standard now includes numerous characters – including, regrettably, several in the Runic range! – that occur only in single historical texts or inscriptions.)

5.2. Option 2: Expand the Runic block to cover all rune-forms

- Add new explicit codepoints for Dalecarlian runes, staveless runes, and all the other observed form variants of existing characters.

This maximalist approach would involve radically expanding the Runic block well beyond its current bounds with dozens of new characters, aiming for much broader coverage of variations in observed runic forms, and encoding them as distinct characters. It goes against Unicode's principle of encoding characters, not glyphs, but in a sense that ship has already sailed (see 4.1. It would also be consistent with the current state of the block, encoding each standardised form separately and explicitly. However, as noted above, this would make text encoding unmanageable for any but the shortest of texts, and would render searching and indexing on the character/grapheme level much more difficult, requiring all form variants to be covered for grapheme-level searches, and complex rules for character-folding and text normalisation. It would, if anything, complicate the kinds of corpus, correspondence and statistical text analyses that scholars may wish to carry out, rather than facilitating them. The authors do not believe that this option would be either desirable or sustainable – it would make an already bad situation worse. A set of standardised forms *is* necessary. But Unicode is explicitly not for encoding glyphs as characters; something we do not question for other scripts.

5.3. Option 3: A minimal character set + markup

- Add missing characters
- But otherwise reduce to a minimal set of distinct characters. (Albeit nonetheless with potentially multiple glyphs at the font level!)

This is the minimalist approach. Removing characters from the existing standard is not possible, and would be highly undesirable even if it were, since there already exist texts encoded using those characters. It may, however, be possible to change character semantics of some characters to declare existing variants deprecated (unlikely) or canonically equivalent for normalisation (more likely, but admittedly not by much). Here we would propose to declare a number of characters representing duplicates and variants of a single grapheme to be canonically equivalent, and one of the pair to be deprecated for use in new texts. Thus through Unicode normalisation rules, the deprecated variants might be normalised to their canonical counterparts. Examples of such pairs include, but are not limited to, the Franks Casket forms (U+16F4—U+16F8), Older Fuþark ᚨ (U+16A8) and Fuþorc ᚫ (U+16AB), and long-branch/short-twig pairs where the forms never represent distinct graph-emes, favouring the existing long-branch character as the canonical one. This approach would require glyph variations (allographs) to be encoded at the markup level, or as rich text, not as separate characters. This has the benefit of preserving reserves graphemic integrity, while allowing for glyph-level encoding, in much the same way as for any other script. However, texts would no longer be able to be accurately encoded as plain-text only. This approach would require runological consensus on a minimal set of graphemes, as opposed to forms, and also a parallel standardisation effort to ensure comprehensive coverage of alternative glyph variants in font support.

Such a minimal character set based upon a subset of the Unicode Runic block has in fact been implemented as a proof of concept: Elisabeth Magin's database of the medieval inscriptions from Bryggen, Bergen [26] has already applied this approach with attribute-based (non-XML) markup, suffixing each runic character with a code specifying the observed form-variant. It's possible to search the database by grapheme, but also by specific form-variants.

5.4. Variation selectors

None of these options are completely satisfactory, with both the maximalist and minimalist approaches forcing compromises in one way or another. However, there may be another way to enable encoding of both the character and the form variant as plain text while keeping the two elements distinct. The Unicode standard itself offers a potential solution, in the form of *variation selectors.*

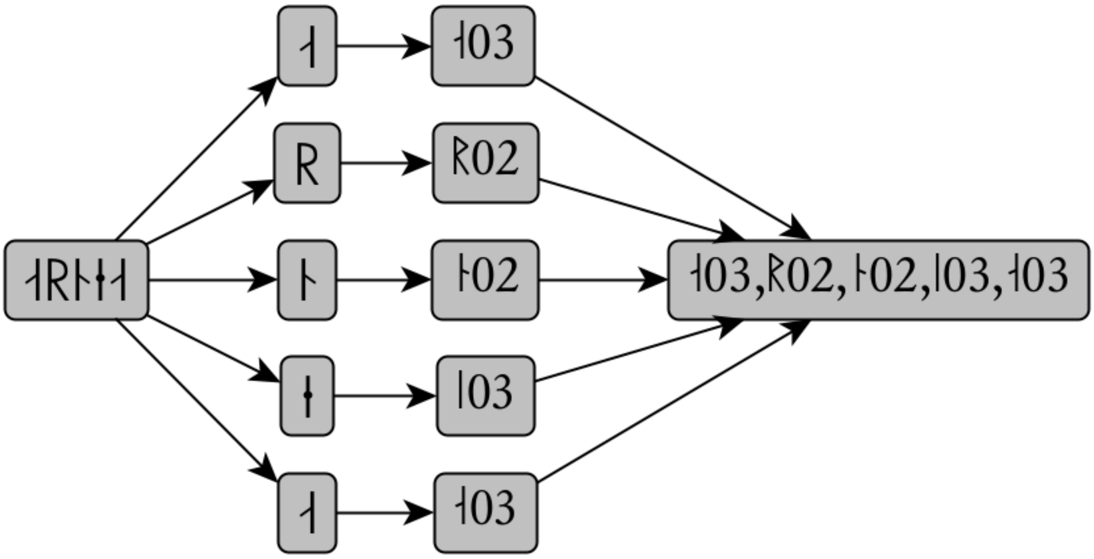


Figure 3: Possible encoding of the inscription from 1 using a minimal character set and attribute-based non-XML markup for graphemic variations of the runes.

Variation selectors are, "non-spacing combing marks. They have no graphic shape of their own; instead they function to pick out a particular, defined subset of potential graphic presentations for the base character to which they are applied" [27]. They provide a way to indicate what form-variant a glyph should take for display, encoded as part of the text. Variation selectors are (relatively) new in Unicode, but fairly widely deployed. They have been used in the Emoji space to distinguish between text and emoji forms of certain characters, for example, as well as to describe historic, obsolete, and other variant forms for Chinese characters. Other applications include specifying a barred variant form for the digit zero. Variation selectors are *modifiers* to the base character. As such, they provide a hint to the rendering engine, that this character should be displayed using a particular form variant, if possible (font support permitting) while not changing the underlying character. Variation selectors degrade gracefully; that is, if font or software support is lacking for a particular form-variant, the correct character will still be displayed, albeit with its "base" or default glyph. To use a hypothetical runic example, you might not be shown a "Greenlandic"-**r** if your font doesn't support it, but you'll still see an **r**. Because the underlying character doesn't change, compliant applications can ignore variation selectors for the purposes of search and indexing; but because the variation selector is still present at the text level, they can also be searched for explicitly as part of a string if desired. This addresses the limitations that would otherwise be concomitant with a minimal character set, and allows for a fourth possible way forward.

5.4.1. Option 4: A minimal character set + variation selectors

- A reduced character set, as Option 3, but instead of allographs at the markup level, define a set of Variation Selectors for each character as part of the Unicode Standard

This would require the same consensus on a minimal set of graphemes, and the same additional standards work to define a set of standard forms as in Option 3 above, but now divided between both the Unicode Consortium for the character/glyph set, and efforts from other bodies for font support for contextual alternate glyphs and stylistic sets for character strings using the variation selectors. Crucially however, the glyph variants would be part of the standard, and no special markup would be required to use and display them – just plain text. Both text encoding and processing, e.g. search, would have the flexibility to work both at the generic character (grapheme) level, *and* at the form variant level, as required. Instead of markup, form-variants could be specified as modifiers to the base character, in plain text, using the appropriate variation selector. Fonts supporting that form would display the appropriate glyph, recognising the string of base-character + variation selector, but fonts without such support would ignore the variation selector and fall back to a generic form for the base character. Similarly, other applications could choose to respect or ignore the variation selector depending on context and the users' intentions.

This would be the authors' preferred solution, and is proposed here as a potentially fruitful avenue for future work. However, it should be noted that despite the existence of Variation Selectors as entities explicitly intended to address the kinds of text encoding needs present in digital runology, the Unicode Consortium have in recent years been increasingly reticent about defining new variation sequences (Harald Tveiten, pers.comm.) and it is possible that any proposal for such sequences attached to characters in the Runic block would be rejected as a matter of policy. This does not mean, however, that variation sequences for runes could not be defined for runological use and data interchange within the domain, outside of but in complement to the Unicode standard.

6. Future work

Runologists, as a user-community, frequently lack familiarity with technology and new digital approaches, and have been more or less inactive with regards to Unicode since the Runic block was published. The fact that very few runologists actively use the Unicode Runic block in their work, almost 25 years after it was created, suggests that it does not meet their needs, and this paper has attempted to explain why, and to propose ways to address those shortcomings. Further inaction and failure to engage with the digital standardisation of the script from working runologists will inevitably lead to one of two outcomes. Either no one else will do anything either, the problem will remain unaddressed, and we *still* won't be able to satisfactorily encode runes; or someone else from outside of runic research will do something and the fate of the script will be out of our hands, guided by the whims of individuals' particular interests – or worse.

With this in mind, the authors propose that runologists, as a user community, discuss the alternative options presented in this paper, and take action. We favour Option 4: a reduced, minimal character set, declaring current duplicates deprecated or canonically equivalent, adding characters that are genuinely not represented, and specifying a set of glyph variants for allographs to be defined as variation sequences. This would require the field to unite in standards work, whether it be in collaboration with or outside Unicode. It would require work within the field to pin-down and reach consensus on a minimal comprehensive set of graphemes, and of allographs. It would then require font support for those glyph variants. Nonetheless, the authors are of the opinion that such work, while potentially challenging and onerous, would greatly benefit the field of runology and allow greater integration with other fields of digital scholarship – including digital scholarship on ancient scripts in general.

References

[1] K. F. Seim, Runologi, in: O. E. Haugen (Ed.), Handbok i norrøn filologi, 2 ed., Fagbokforlaget, Bergen, 2013, pp. 128–193.

[2] M. P. Barnes, Runes: a handbook, Boydell Press, Woodbridge, 2012.

[3] T. Spurkland, Norwegian runes and runic inscriptions, Boydell & Brewer, 2005.

[4] W. Lundström, Proposal to ISO/IEC JTC1/SC2/WG2 Concerning Inclusion into ISO/IEC 10646 of the Repertoire of Runic Characters, Technical Report, 1995-01-08.

[5] T. Spurkland, Kriteriene for datering av norske runesteiner fra vikingtid og tidlig middelalder, Maal og Minne (1995) 1–14.

[6] A. Palumbo, Skriftsystem i förändring: En grafematisk studie av de svenska medeltida runinskrifterna, Uppsala: Institutionen för nordiska språk vid Uppsala universitet, 2020. URL: http://www.diva-portal.org/smash/record.jsf?pid=diva2:1509013&dswid=-3719.

[7] M. P. Barnes, Standardised fuþarks: A useful tool or a delusion?, Arkiv för nordisk filologi (2006) 5–22. URL: https://journals.lub.lu.se/anf/article/view/11743/10422.

[8] M. Schulte, The rise of the younger fuþark, NOWELE 60-61 (2011) 45–68. doi:10.1075/nowele.60-61.03sch.

[9] K. F. Seim, De vestnordiske futhark-innskriftene fra vikingtid og middelalder – form og funksjon, Ph.D. thesis, NTNU Trondheim. Det historisk-filosofiske fakultet. Institutt for nordistikk og litteraturvitenskap, 1998.

[10] S. Fridell, Graphic variation and change in the younger fuþark, NOWELE 60-61 (2011) 69–88. doi:https://doi.org/10.1075/nowele.60-61.04fri.

[11] M. P. Barnes, On the status and transliteration of the additional characters of medieval scandinavian runic writing, Amsterdamer Beiträge zur älteren Germanistik 67 (2011) 9–21. doi:10.1163/9789401200783_003.

[12] K. Düwel, Runenkunde, 4 ed., Metzler, Stuttgart, 2008.

[13] A. Liestøl, I. S. Johnsen, J. E. Knirk, Norges innskrifter med de yngre runer: Bryggen i Bergen, number VI in Norges innskrifter med de yngre runer, Kjeldeskriftfondet, Oslo, 1980-1990.

[14] M. Waldispühl, Schreibpraktiken und Schriftwissen in südgermanischen Runeninschriften, number 26 in Medienwandel - Medienwechsel - Medienwissen, Chronos, 2013.

[15] S. Nowak, Schrift auf den Goldbrakteaten der Völkerwanderungszeit: Untersuchungen zu den Formen der Schriftzeichen und zu formalen und inhaltlichen Aspekten der Inschriften = Writing on the Migration-Period Gold Bracteates, Ph.D. thesis, Georg-August-Universität zu Göttingen, Philosophische Fakultät, 2003. URL: http://resolver.sub.uni-goettingen.de/purl/?webdoc-512.

[16] T. Spurkland, En fonografematisk analyse av runematerialet fra Bryggen i Bergen, Ph.D.

thesis, Institutt for nordistikk og litteraturvitenskap, Institutt for arkeologi, kunsthistorie og numismatikk, Universitetet i Oslo, 1991.

[17] L. Peterson, Scandinavian runic-text data base: a presentation, in: B. Ambrosiani, H. Clarke (Eds.), Developments around the Baltic and the North Sea in the Viking Age, volume 3 of *Birka studies*, Viking Congress, Birka project for Riksantikvarieämbetet and Statens Historiska Museer, Stockholm, 1994, pp. 305–309.

[18] H. Gustavson, S. Jörsäter, Runes and the computer, in: C. W. Thompson (Ed.), Proceedings of the First International Symposium on Runes and Runic Inscriptions, number VII in Michigan Germanic Studies, Department of Germanic Languages and Literatures, The University of Michigan, Ann Arbor, Michigan, 1981, pp. 98–106.

[19] C. W. Thompson, On transcribing runic inscriptions, in: C. W. Thompson (Ed.), Proceedings of the First International Symposium on Runes and Runic Inscriptions, number VII in Michigan Germanic Studies, Department of Germanic Languages and Literatures, The University of Michigan, Ann Arbor, Michigan, 1981, pp. 89–97.

[20] A. Doyle, Digital ogam: Implementation and implications, 2022. URL: https://ogham.glasgow.ac.uk/index.php/2022/05/23/digital-ogam-implementation-and-implications-guest-blog-by-adrian-doyle/.

[21] A. West, Babelstone: Khitan scripts, ???? URL: https://babelstone.co.uk/Khitan/index.html.

[22] The Unicode Consortium, J. Becker, M. Davis, M. Everson, A. Freytag, J. Jenkins, E. Muller, L. Moore, M. Suignard, K. Whistler, The Unicode Standard, Version 4.0, Addison-Wesley Professional, 2003. URL: https://www.unicode.org/versions/Unicode4.0.0/ch13.pdf.

[23] M. Deroń, Proposed additions to the runic range, l2/09-312, 2009. URL: http://unicode.org/L2/L2009/09312r-runic-additions.pdf.

[24] The unicode standard, version 7.0, runic range: 16a0–16ff, 2014.

[25] J. D. Becker, Unicode 88 (standard), 1988. URL: https://www.unicode.org/history/unicode88.pdf.

[26] E. M. Magin, Runes, Runic Writing and Runic Inscriptions as Primary Sources for Town Development in Medieval Bergen, Norway, Ph.D. thesis, University of Nottingham, School of English, Nottingham, 2021.

[27] Glossary of unicode terms, 2023. URL: https://www.unicode.org/glossary/.

The Digital Lab as an arena for teaching and outreach activities connected to the Special Collections at the University Library of Bergen

Emma Josefin Ölander Aadland

The University of Bergen Library, Haakon Sheteligs plass 7, 5007 Bergen, Norway

Abstract

The Digital Lab at the University of Bergen Library was established in 2020 and is set up to be an interdisciplinary hub for researchers, lecturers, and students, both on-site and digitally. The lab provides a space to learn, to discuss, and to apply various digital tools and methods used in research within the Digital Humanities. Each semester the lab sets up different courses, workshops, seminars, and lectures aiming to support and serve the target groups.

This paper presents a case study that investigates how the Digital Lab can provide an arena for DH activities connected to the Special Collections at the library. The aim is to explore how these activities can have a cross-disciplinary and collaborative approach, and to investigate the lab's potential as a community-building arena and its impact on the involvement of the library and the librarian‘s role in DH.

Keywords

Cross-disciplinary, Digital Humanities Lab, Community building, Library, Librarian

1. Introduction

In recent years, as the field of Digital Humanities (DH) has evolved, many libraries have felt the need to explore the role of the library and librarian in relation to DH. Some libraries have responded by establishing initiatives such as labs, hubs, centers, or Maker Spaces, aiming to provide primarily technical, methodological, and infrastructural support for DH scholarship [1]. Another approach adopted by libraries and librarians in relation to DH has been to act as supportive contributors in outreach and teaching activities [2, p. 239]. Additionally, certain libraries have participated in various DH projects, while others have played a central role in community collaboration [2, p. 157]. Regardless of the level of involvement or chosen approach, it can be argued that DH has somewhat challenged the role of both libraries and librarians. However, the library's involvement in DH research or other collaborative connections to DH are not extensively addressed in DH literature [3, p. 139]. This paper aims to explore how the involvement of libraries and librarians in DH offers both possibilities and challenges that can contribute to reshaping or redefining the librarian's role.

DHNB2023 | Sustainability: Environment - Community - Data. The 7th Digital Humanities in the Nordic and Baltic Countries Conference. Oslo – Stavanger – Bergen, Norway. March 8–10, 2023.

emma.aadland@uib.no (E. J. Aadland)

DHNB Publications, DHNB2023 Conference Proceedings, https://journals.uio.no/dhnbpub/issue/view/875

The aim is to examine how various Digital Humanities (DH) activities in the Digital Lab at the University of Bergen Library, employing a cross-disciplinary and collaborative approach, can contribute to community-building and simultaneously develop the role of the library and librarians in the context of DH. The approach involves using these DH activities as a case study to investigate how the lab can serve as an arena for both teaching and outreach activities involving the Special Collections at the library. The outreach activities primarily consist of seminars associated with the library's exhibitions, showcasing different components of the University's Special Collections in both physical and digital formats. On the other hand, the teaching activities encompass courses and workshops conducted in collaboration with the staff responsible for the collections.

The hypothesis is that the Digital Lab can actively participate in existing academic activities within the library and various academic communities, serving as a sustainable cross-disciplinary space for students, researchers, and lecturers. Simultaneously, the lab can play a significant role in facilitating teaching and outreach activities that make the Special Collections accessible to a broader audience and furthermore foster research opportunities. To investigate the lab's potential as a community-building arena and its impact on the involvement of the library and librarians in DH, the following questions are addressed:

1. *What challenges and opportunities do DH activities connected to the Special Collections in the library present to the Digital Lab for community-building?*
2. *In what ways does the library's engagement in DH activities challenge the role of librarians?*

2. The case study

To address the questions, the following sections will provide an overview of both the Digital Lab and the Special Collections at the University of Bergen Library, along with a detailed description of the outreach and teaching activities associated with them.

2.1. The Digital Lab

The Digital Lab at the University of Bergen Library was established in 2020 and is set up to be an interdisciplinary hub for researchers, lecturers, and students, both on-site and digitally. The lab provides a space to learn, to discuss, and to apply various digital tools and methods. Each semester the Digital Lab hosts different courses, workshops, seminars, and lectures aiming to support and serve the target groups.

A working group consisting of five staff members at the library, with different technical and disciplinary backgrounds, has the operative responsibility for the lab. Other library staff members contribute to the lab‘s activities, for example by offering expertise in teaching tools or sharing experience as a supportive partner involved in various collaborative DH projects.

The lab seeks collaboration with different academic communities, and for example coordinates the Digital Humanities Network at the University of Bergen. The Network involves researchers, PhD candidates, postdocs, and staff mainly from the Humanities and Social Science. An advisory board for the network is established with the aim of keeping the lab's activities relevant to the members.

Another collaborative and cross-disciplinary example is the Collaborative Scientific Software Development Summer School for PhD candidates and postdocs from different disciplines, first time conducted in June 2022 in cooperation with the Department of Informatics at the University of Bergen. In 2023 this collaboration has been extended to include the discipline of Digital Culture at the university, with the aim of increasing engagement from the Humanities and Social Science and fostering even more cross-disciplinary collaboration.

2.2. The Special Collections and the exhibitions

The Special Collections at the University of Bergen Library consist of the Language Collections, the Picture Collection, the Manuscripts and Rare Books Collection, and the Queer Archive. These collections held both digital and digitized material as well as physical artifacts.

The first exhibition associated with the Special Collections was established in the autumn of 2021 within the Arts and Humanities library at the University of Bergen. This exhibition space was created through library renovations and updates specifically designed to accommodate exhibitions. The exhibitions showcase primary sources, digital materials, and digitized items from various archives within the Special Collections. Typically, the exhibitions run for approximately four to six months. Some exhibitions also have a digital format, and the project group responsible for the library exhibitions collaborates with different academic communities at the university.

2.3. The outreach activities in connection with the exhibitions

The Digital Lab, in collaboration with the Special Collections, has developed a concept where contributors who have been involved in the library exhibitions, such as curators or professional advisors, as well as researchers working on topics related to the exhibition theme, are invited to the lab for seminars. The purpose of these seminars is to provide the audience with insights into how the contributors have engaged with primary sources, digital materials, and digitized items in relation to the exhibition and their own research. Additionally, the seminars serve as a platform for discussing critical questions raised in the exhibition.

The first seminar, held in November 2021, was a brief and general introduction to the different parts of the Special Collections, with the intention to continue with a seminar series focusing on each of the collections separately. This developed into the breakfast seminar series in connection to the exhibitions in the library, and the first breakfast seminar was held in August 2022. After that, one seminar was held in December of the same year, and two in the first quarter of 2023.

The target groups for these seminars, including students, academic staff, and researchers, have been partially reached based on the participants who registered for the seminars. On average, around 20 participants attended each seminar, representing a diverse range of faculties and fields of study.

2.4. Teaching activities in connection with the Special Collections

The Digital Lab has been an arena for developing various activities, including workshops and courses in which the Special Collections at the library are involved to some extent. The main part of these activities is developed in collaboration with the Language Collections. An example here is “The introductory course to the use of the language data infrastructure CLARIN”. CLARIN is a digital infrastructure offering data, tools and services to support research based on language resources [4], and the course targets both students and researchers. It has a research support approach including sections about, for example, depositing data sets and the FAIR principles [5]. The course primarily attracts students studying language, linguistics, or related fields, limiting its potential for cross-disciplinary engagement.

Another example is "The Corpus Workshop”, which explores corpus tools and resources provided by the Language Collections in the language research infrastructure CLARINO [6]. The workshop provides examples of data extraction suitable for research and assignments, comprising introductory lectures and hands-on exercises. The instructors involved include both library staff and staff from the Department of Linguistic, Literary and Aesthetic Studies at the University of Bergen.

The Digital Lab also organizes workshops and courses with a more cross-disciplinary approach with the use of familiar DH tools, such as the Geographic Information System (GIS), which is a tool that creates, manages, analyzes, and maps all types of data. The GIS workshops are held regularly, and have attracted both students, young researchers, and staff from different fields within the university. Lastly, there is "The Course in Copyright for Photographs," involving an instructor from the Picture Collection. This course addresses common issues encountered in image research and the daily usage of images.

3. Discussion

The raised questions in the introduction pertain to the challenges and possibilities presented by DH activities related to the Special Collections in the library for community-building. Additionally, they explore how these activities can potentially challenge the roles of the library and librarians in the context of DH. The subsequent sections will delve into these aspects from various perspectives.

3.1. The library and librarian’s role in Digital Humanities

The library has for a long time been concerned with facilitating, organizing, providing access, and sustainable preservation for both primary sources in traditional forms, and materials in digital or digitized formats. Here, the library and Digital Humanities share interest when it comes to collection, organization, preservation, and use of digital materials [3, p. 137]. However, the focus often is on whether the library has a role in DH. It can therefore be proposed to shift this focus and instead ask how the library can contribute to DH, because it is pertinent for the library to engage with the field when it comes to achieving the goal of providing accessibility of information. Both the library and DH aim to increase the digital accessibility and research potential of cultural materials [7]. As the university library has an initial function as an interdisciplinary agent for the different faculties at the university, the library is overall in many ways obliged to respond to the different needs of all of them. The support given by the library is therefore comprehensive, and regarding DH support it is argued that librarians should be encouraged to investigate their role in DH more widely [7].

Sula [8, p. 24] asserts that libraries are well positioned to meet complex needs in relation to DH scholarship. However, studies show that librarians experience that their expertise is undervalued or instrumentalized as a part of a service model that all together separates the librarians from an academic level of inference in the DH field [9]. Librarians may be claimed to be generalists in DH, as they consult on a range of methods across disciplines. At the same time, it is argued that librarians face a diversity of challenges in connection to DH and that it must be ensured that they learn the requisite skills and have the essential knowledge and experience to meet the needs [10, p. 135]; [11]; [1, p. 365]. This can be challenging for librarians working in connection to or directly in the DH-field, as it is a constantly evolving field. In other words, this raises some crucial questions about the challenges the staff in the library, who are involved in DH, meet regarding user needs and how these can be addressed.

The DH support provided by staff at the library has for some time primarily been technical services and the fostering of collection accessibility and visibility. Burns [2, p. 239], on the other hand, points out that a part of the librarian's role has developed more into connecting students, researchers and teachers to relevant information and to be a collaborator in the activities connected with research and educational environments. Universities providing exclusive study programs or introductory courses in DH are becoming more common at many universities [12]. At the same time, an established practice of bringing libraries directly into DH courses provided by universities is still somewhat unusual [12, p. 331]. The connection between DH, libraries and librarians can therefore be described as topical, and the absence of librarians in the classroom where DH is taught is to some extent deficient [12, p. 343].

On the other hand, the development for some libraries has been heading in a direction that has solidified their position and role in the field of DH [13, p. 5]. In connection with this, there has been a growing emphasis on the development of training tools and teaching methods to support software development and infrastructure requirements [14, p. 159]; [1]. Gooding [3, p. 143] states that this has made scholars more avidly interested in the support and contribution that libraries can give regarding DH, and he argues that the link between libraries and DH can be understood as convergence points between different disciplines [3]. During the past two decades it has also become more common to place physical DH centers, hubs or labs in university libraries, and this points out the library as a provider of DH support both related to research and teaching [15]; [3, p. 138]. Initiatives like these can therefore be considered as an answer to the rapid growth of the DH field.

Furthermore, the role the library may have in DH projects depends on a range of factors. For some research libraries the role as a project owner is obvious, while there are strings attached to the involvement for other libraries that limit the role to being a partner, contributor or collaborator. How the library and librarians engage with the field of DH also differs to some extent, depending on how well the engagement is supported both by library leaders and the leaders of the faculties and the university (see e.g., [16]). Some libraries and librarians also experience challenges connected to new forms for collaboration between faculties and the library concerning the switch from the library being first and foremost considered as a service provider to the development of being a collaborator or partner in DH [10].

For the University of Bergen Library, the DH involvement of the library staff has differed. As mentioned, many of the staff members, both academic librarians, senior advisors and developers, are involved in the activities in the Digital Lab, as instructors or contributors of DH support in other ways. Before the lab was established, however, the engagement in DH was as a project partner or collaborator in DH projects owned by different faculties, and for the library staff this mainly concerned technical contributions such as developing and building technical infrastructure and giving sustainable long-term preservation. This involvement therefore primarily included the staff working as developers at the library and did not engage the librarians to a particularly large extent.

It is therefore thinkable that the establishment of the Digital Lab has been a contributor to extend the librarians‘ involvement in different outreach and teaching activities in connection to DH. As the establishment of the lab is a direct answer to the need of technical and scholarly support outlined in the Humanities Strategy for the University of Bergen, this also underlines the importance of the library's adaptation to changes in needs from the different academic communities. The rapid changes both for the library and the staff at the library unarguable also bring both challenges and possibilities regarding for example taking part in community-building with the Digital Lab as an arena.

3.2. The library and librarians‘ role in community-building

As to community-building, there are some challenges that can be addressed. It can, to begin with, be time demanding and require commitment [17, p. 26]. Therefore, it is essential to put in an effort and to have a structure around community building and committed collaboration that can benefit from involving the library, for example through the initiative of a lab. As mentioned, The Digital Lab coordinates the DH Network at the University of Bergen. This network is to some extent an established DH community, but it could in many ways benefit from a more interdisciplinary compound with more members from other faculties than the Humanities and Social Sciences. To accomplish that other disciplines find it relevant to be a part of the community, it is relevant to enable several regular and various activities in the lab. A range of activities such as project meetings, lectures, seminars, training sessions, or informal get-togethers, may all together turn out to create a sense of community [13, p. 17]. In addition, much learning is done when members of a DH community share their work, expertise and experience [13, p. 35].

However, it can also be perceived as a challenge to build and maintain a community of practitioners with diverse backgrounds and skills, to meet the different needs and to be relevant for all in the community. It can therefore be crucial to invest more in building and maintaining relationships with DH practitioners in order to support the growth and sustainability of the field. To meet the different needs of the researchers in the DH Network, it is of course important to ensure that the DH activities in the lab are varying and of academic relevance. One way is to improve the engagement in the network in terms of participation in the activities hosted by the lab as well as the collaboration around these activities. Furthermore, this can be an argument regarding the librarians' involvement in DH scholarship.

Another challenge regarding community-building has to do with the researchers' own identity as a DH scholar, and many scholars would not label themselves as such [15, p. 13, 33]. Someone's engagement in DH should not, however, be determined based on if the person in question identifies as a digital humanist. This is relevant to be aware of when it comes to DH being a field for interdisciplinary research, or as Svensson [15] states: "what is important is that scholars and experts across a range of disciplines and specialties come together and contribute to humanities-driven exploration of digitally inflected research and education". At the same time, an awareness of what Bell and Kennan [18, p. 166] point out should not be forgotten: "to gain equal recognition as digital humanists, librarians must step outside the library and embrace being digital humanists themselves". In other words, the exploration of collaboration in connection with community-building also depends on the approach of the librarians wanting to engage with DH. And furthermore, how this involvement is supported by the library leaders.

A relevant question is also how familiar DH as a field is for students. Outreach activities, like the breakfast seminars tried out in the Digital Lab, are one way to make DH methods, tools and research more visible for students from different disciplines as well as position the library as a suitable arena for DH activities. However, the success of these outreach activities somewhat depends on the actual participants in the seminars and if one has reached out to the target groups aimed for. It can also be significant, in an extension of these seminars, to explore how the activities can be followed up by other events like workshops or courses, where students and researchers can come together and explore the possibilities for research based on sources from the Special Collections.

Another way to build a community for DH scholars is to ensure that the library is involved in developing more DH projects, and furthermore to ensure that at least some of these projects involve the Special Collections at the library. The Digital Lab could potentially use DH projects as an opportunity to engage with the public through different outreach activities and in this way, both promote the value and impact of DH research and expose different parts of the collections. At the same time, this underlines the need for institutional support for libraries seeking to establish a role and presence in DH [17, p. 26]. The dialogue between the library and the faculties regarding the library's role in DH projects can benefit from an established framework that outlines the different parts of both the library and librarian's involvement.

3.3. Challenges and possibilities for a sustainable involvement in DH

Recent Library and Information Science (LIS) literature shows skepticism towards the traditional service models for libraries and librarians, and many suggest a turn away from mainly support and service, and instead towards librarians being collaborative partners [18, p. 164]; [16]. In an extension of this, a relevant question for many libraries is if the service and support provided solely supports DH scholarship or if the library and librarians also are co-producers of the DH research.

The University of Bergen Library's role in DH projects is to some extent limited when it comes to being a project owner, so for now the possibility for involvement is mainly by ensuring that one seeks partnership with, for example, faculties or other institutions in DH projects. Over the last decade the number of research projects involving the library as a collaborative partner has in fact increased and, as pointed out earlier, the involvement of the library has so far mainly been technically related to infrastructural support or taking part as project coordinator. In other words, there is a potential to explore what the library's role can develop into. One turn, that at the same time can promote sustainable operation for the Digital Lab, is to take an extensive role in both outreach and educational activities as a part of the DH projects. It is therefore crucial that the lab is considered as a suitable arena for these activities and that one ensures that necessary resources at the library have available capacity to enable the involvement of the library staff.

As already mentioned, DH courses and educational activities for students are increasing at many universities [12]. This provides an opportunity for involvement in DH as an increasing DH support regards teaching and student learning. The library can contribute by offering training and support in tools and exploring ways to use technology more creatively both in teaching and research. It therefore is crucial that librarians seek collaboration and partnership to gain the potential for learning outcomes across disciplines at campus [1, p. 372]. Furthermore, initiatives like labs, centers, hubs or similar might be a beneficial approach for the library to create a suitable environment for DH activities at the university. These initiatives serve as attractive arenas for scholars from diverse disciplines to engage in DH scholarship. The possibility to create a sense of DH community might also be easier when the library is involved by providing a suitable meeting place for community activities. Here, the link between libraries and DH can be understood as convergence points between different disciplines, which makes the library a prominent space for cross-disciplinary collaborations.

The University of Bergen does not offer a study program in Digital Humanities, but currently offers two MA level courses on specific methods in the digital humanities in the disciplines of Digital Culture and Archeology. However, an increasing number of disciplines are engaging with either DH tools or methods, or both, and therefore it is a need for DH educational activities at the university. This should therefore be an argument of relevance for the library regarding even more engagement in DH activities because the gap to fill can be solved by involving both the librarians and the lab to a greater extent. In advance, this may turn out to be a way for both the library and librarians to establish a role in relation to DH. In this regard, involving librarians who serve as Subject Specialists in teaching activities within the lab could offer a sustainable solution. These librarians often have direct connections with academic communities and possess extensive experience in teaching various library courses. Many are also involved in other academic activities and therefore have established contact with academic communities and an insight into the specific needs for DH support.

Svensson [15, p. 17] argues that it is necessary that scholar work has an awareness of the need of infrastructure and methodological competence, but also that infrastructure works lack something important if they are not connected to exciting scholarly and archival challenges. He exemplifies that different disciplines and scholarly traditions have different ways of engaging with the digital. Therefore, it can be crucial that the Digital Lab continues in using teaching and outreach activities to demonstrate the value and impact of digital technologies for both students and researchers. Furthermore, the resources in form of both tools and databases as well as the primary sources and digital or digitized sources from the Special Collections should be made more accessible for students and researchers to generate more research based on these sources. The breakfast seminar series in connection with the exhibitions may be seen as a first step, but these activities are surely not enough if they do not result in more scholarly-based activities. To gain full success with this type of activity, it might be necessary to follow up with other DH activities for students and researchers that result in dissertations and research.

3.4. Common DH values as a guideline

Both shared values and spaces can have an impact on building partnerships between academics and librarians, and in advance also make DH and the librarian's role in the field more visible [18]. Openness, collaboration, collegiality and connectedness, diversity, and experimentation are common DH values [19] that the library and librarians wanting to get involved with DH can promote. Even though these values cannot be expected to gain full consensus [19, p. 19], they may have a significant role for the library wanting to establish a suitable space for facilitating cross-disciplinary methods. Or as Vandegrift [7] puts it: "The library must function as a place where scholars can try new things, explore new methodologies and generally experiment with new ways of doing scholarship".

To enable the potential of interdisciplinary research within DH, an explorative approach can be crucial. At the same time, one should ensure that there is an arena for this exploration. To create a culture of openness and inclusivity in initiatives like labs, hubs, centers and Maker Spaces, it is important that both students, researchers and staff from diverse disciplines feel welcomed and valued. This can also contribute to creating a sense of community and encourage cross-disciplinary collaboration. Furthermore, a mixture of different DH activities might be the way of embracing diversity in the community. The library can also benefit from a deeper involvement in academic activities related to DH and make sure that librarians and other staff members at the library with connections to different disciplines are involved. This can help to encourage researchers to collaborate across disciplines and to see the value of working together and sharing their knowledge and expertise.

Furthermore, in order to facilitate cross-disciplinary collaboration, it is important to encourage researchers to work together and share their knowledge and expertise. This can be done through regular meetings, workshops, and other opportunities for researchers to come together and present or discuss their work. In connection with this, it can be underlined that librarians may be seen as "interdisciplinary mediators" suitable for meeting needs in any discipline [18, p. 167]. Once more it can be underlined that the connection between librarians and other academic staff is necessary and should be ensured through different activities where the lab provides a suitable arena or meeting point.

It is crucial, however, to be aware that researchers from different disciplines may find it challenging to work together, especially if they are not familiar with each other's methods and approaches [3, p. 149]. Providing training and support for cross-disciplinary research can help to bridge these gaps and facilitate collaboration and connectedness. The library with a lab or a similar initiative can provide the eminent space whereas the staff at the library can take part in offering the training and support needed. As to teaching activities, these may promote the use of digital technologies in research and can accentuate the potential of these technologies for advancing knowledge and understanding.

The relevance of highlighting the value of cross-disciplinary research and its potential to generate new insights and understanding through experimentation is necessary in this context. To enable this, it is important to have a space aimed for exploration and embrace the diversity of the different research fields. This can help to encourage researchers to collaborate across disciplines and to see the benefits of working together. The library with the Digital Lab can serve as a provider of a physical and digital space, where scholars from different fields can connect, experiment and get involved with DH. In an extension, the lab may explore new models for sustainable operation, such as partnerships with other organizations or collaborations where the development of digital platforms and tools can be explored and hopefully used by a wide range of researchers and scholars. The value of collaboration also concerns building relationships with other DH labs and hubs, as this can help to facilitate cross-disciplinary collaboration and provide opportunities for researchers to work with colleagues from other institutions. This can for example be done through exchanges, joint projects, and other forms of collaborations where also the library with its staff is involved.

4. Conclusion

The library and librarians' involvement in DH challenges to some extent the role of both. For the library it can be challenging to response to and meet the different needs of a diversity of academic communities and to make a turn from a service provider for the faculties to a contributor in DH research and educational activities [10]. For librarians, on the other hand, it can be challenging to ensure that one has the requisite skills and the essential knowledge and experience to meet the needs in DH as it is a field in constant development [10, p. 135].

For some libraries the engagement in DH has encouraged an explorative approach with a result of involvement both in DH projects as well as different DH activities such as teaching, training and research. For many, the establishment of initiatives like labs, hubs, Maker Space and centers has been a way of engaging with DH and ensuring that the library provides an eminent space for research, outreach and teaching activities.

However, the connection between DH, libraries and librarians is still to some extent topical, and the lack of involvement in teaching activities is deficient. Hence, it is crucial not to underestimate the significance of adopting a perspective where the library considers its involvement in DH as inherent. To effectively engage in DH endeavors, it can be beneficial for the library to embrace key values commonly associated with DH, such as openness, collaboration, collegiality, connectedness, diversity, and experimentation. By embracing these values, the library can foster an environment conducive to DH scholarship and effectively contribute to the advancement of the field [19].

At the same time, one crucial factor that might need to be established to enable cross-disciplinary collaboration that involves the library and librarians is the sense of a DH community. A successful and operative DH community, however, might require a better understanding of the role of both the librarian as a digital humanist and the researcher as a DH scholar. Furthermore, one should not underestimate the value of a suitable space for different activities that can contribute to building a community of DH scholars. Here, the DH values may promote a more coherent identity for the DH community and its members [19, p. 30].

As this case study shows, the different DH activities in connection with the Special Collections has been a way for the Digital Lab in the library to reach out both to researchers and students from across disciplines. The teaching activities in collaboration with the Special Collections may play a significant role in enabling exploration of both technologies and methodologies, while the outreach activities are a way to encourage exploration regarding ways to use sources from the collections as subject for research. These DH activities in the lab encourage cross-disciplinary collaborations but have not yet resulted in any concrete DH project or research that involves the library. Therefore, it can be argued that the actual interdisciplinarity also depends on the approach of the scholars from the different disciplines and to what extent disciplinary boundaries possibly can prevent interdisciplinary collaborations and research [3, p. 149]. In other words, the DH activities in the library may be a vital facilitative factor, but the outcome of these activities and the actual possibility of interaction between different fields or disciplines also depends on each scholar engaging in the activities.

References

[1] J. Nichols, M. M. Melo, J. Dewland, Unifying Space and Service for Makers, Entrepreneurs, and Digital Scholars, Libraries and the Academy 17 (2017) 363–374. doi:10.1353/pla.2017.0022.

[2] J. A. Burns, Role of the Information Professional in the Development and Promotion of Digital Humanities Content for Research, Teaching, and Learning in the Modern Academic Library: An Irish Case Study, New Review of Academic Librarianship 22 (2016) 238–248. doi:10.1080/13614533.2016.1191520.

[3] P. Gooding, The Library in Digital Humanities: Interdisciplinary Approaches to Digital Materials, in: K. Schuster, S. Dunn (Eds.), Routledge Handbook on Research Methods in Digital Humanities, Taylor & Francis eBooks Complete, Routledge: Oxon, UK, 2020, pp. 137–151.

[4] C. ERIC, The Research Infrastructure for Language as Social and Cultural Data, n.d. URL: https://www.clarin.eu/.

[5] G. FAIR, FAIR Principles, n.d. URL: https://www.go-fair.org/fair-principles/.

[6] CLARINO, The CLARINO Bergen Centre, n.d. URL: https://repo.clarino.uib.no/xmlui/.

[7] M. Vandegrift, What Is Digital Humanities and What's it Doing in the Library?, The Library with the Lead Pipe (2012). URL: https://www.inthelibrarywiththeleadpipe.org/2012/dhandthelib/.

[8] C. A. Sula, Digital Humanities and Libraries: A Conceptual Model, Journal of Library Administration 53 (2013) 10–26. doi:10.1080/01930826.2013.756680.

[9] C. A. Sula, S. E. Hackney, P. Cunningham, A Survey of Digital Humanities Programs, The Journal of Interactive Technology and Pedagogy 11 (2017). URL: https://jitp.commons.gc.cuny.edu/a-survey-of-digital-humanities-programs/.

[10] C. Millson-Martula, K. Gunn, The Digital Humanities: Implications for Librarians, Libraries, and Librarianship, College & Undergraduate Libraries 24 (2017) 135–139. doi:10.1080/10691316.2017.1387011.

[11] M. D. Poremski, Evaluating the Landscape of Digital Humanities Librarianship, College & Undergraduate Libraries 24 (2017) 140–154. doi:10.1080/10691316.2017.1325721.

[12] K. Mapes, Discovering Digital Humanities Methods Through Pedagogy, in: K. Schuster, S. Dunn (Eds.), Routledge Handbook on Research Methods in Digital Humanities, Taylor & Francis eBooks Complete, Routledge: Oxon, UK, 2020, pp. 331–352.

[13] V. Lewis, L. Spiro, X. Wang, J. E. Cawthorne, Building Expertise to Support Digital Scholarship: A Global Perspective, 2015. URL: https://www.clir.org/pubs/reports/pub168/.

[14] J. Smithies, A. Ciula, Humans in the Loop. Epistemology and Method in King‘s Digital Lab, in: K. Schuster, S. Dunn (Eds.), Routledge Handbook on Research Methods in Digital Humanities, Taylor & Francis eBooks Complete, Routledge: Oxon, UK, 2020, pp. 155–172.

[15] P. Svensson, Introducing the Digital Humanities, in: Big Digital Humanities: Imagining a Meeting Place for the Humanities and the Digital, University of Michigan Press, 2016, pp. 1–35. URL: http://www.jstor.org/stable/j.ctv65sx0t.5.

[16] M. P. Posner, No Half Measures: Overcoming Common Challenges to Doing Digital Humanities in the Library, Journal of Library Administration 53 (2013) 43–52. doi:10.1080/01930826.2013.756694.

[17] L. Wilms, C. Derven, L. O'Dwyer, K. Lingstadt, D. Verbeke, M. Lefferts, Europe's Digital Humanities Landscape: A Study From LIBER's Digital Humanities & Digital Cultural Heritage Working Group, 2019. doi:10.5281/zenodo.3247286.

[18] E. C. Bell, M. A. Kennan, Partnering in Knowledge Production: Roles for Librarians in the Digital Humanities, Journal of the Australian Library and Information Association 70 (2021) 157–176. doi:10.1080/24750158.2021.1907886.

[19] L. Spiro, ‘This is Why We Fight‘: Defining the Values of the Digital Humanities, in: M. K. Gold (Ed.), Debates in the Digital Humanities, University of Minnesota Press, Ebook Central (ebrary), 2012, pp. 16–35.

Transliteration Model for Egyptian Words

Heidi Jauhiainen[1], Tommi Jauhiainen[1]

[1]*Department of Digital Humanities, University of Helsinki, Helsinki, Finland*

Abstract

In this paper, we describe token-based transliteration models for Egyptian words. We explain how we created them using an automatic alignment method we devised based on the Needleman-Wunsch sequence alignment algorithm. We use two sources where encoded Egyptian hieroglyphs and their transliteration pairs are available. Ancient Egyptian Sentences (AES) includes a collection of texts where c. 254,000 Egyptian words encoded using Manual de Codage (MdC) have been aligned with their transliteration counterparts. The second source is the Ramses Transliteration Corpus (RTC), with almost 500,000 MdC encoded words. The RTC consists of encoded hieroglyphic sentences, each on its line, and respective transliteration lines in another file. However, unlike the AES, there is no ready alignment of the MdC and its transliteration on the word level. In order to find word-transliteration pairs, we align the sentences of encoded words with the respective transliterations. The alignment task is made more difficult because many of the texts contain damaged parts and editorial additions.

Keywords

language modeling, word alignment, transliteration, hieroglyphs, Ancient Egyptian

1. Introduction

In this paper, we describe our current work on creating token-based transliteration models for Egyptian words, which are needed for an automatic transliteration method we aim at developing. In order to create such models for transliteration, a corpus of machine-readable Egyptian hieroglyphic texts with their transliterations is needed. We have identified two sources where encoded Egyptian hieroglyphic words and their transliteration pairs are available. *AES - Ancient Egyptian Sentences*, based on texts from the Thesaurus Linguae Aegyptiae (TLA) [1], includes a collection of sentences where Egyptian words encoded using Manual de Codage (MdC) have been aligned with their transliteration counterparts [2]. The second, even more extensive, source of MdC encoded and transliterated words is the *Ramses Transliteration Corpus* (RTC) [3]. MdC is the most often used encoding for hieroglyphic texts [4, 5, 6]. The RTC consists of encoded hieroglyphic sentences, each on its own line, and respective transliteration lines in another file. However, unlike the AES, there is no ready alignment of the MdC and its transliteration on the word level. In order to find word-transliteration pairs, we need to align the sentences of encoded words with the respective transliterations. The alignment task

DHNB2023 | Sustainability: Environment - Community - Data. The 7th Digital Humanities in the Nordic and Baltic Countries Conference. Oslo – Stavanger – Bergen, Norway. March 8–10, 2023.

{firstname.lastname}@helsinki.fi (H. Jauhiainen); {firstname.lastname}@helsinki.fi (T. Jauhiainen)
https://researchportal.helsinki.fi/en/persons/heidi-jauhiainen (H. Jauhiainen);
https://researchportal.helsinki.fi/en/persons/tommi-jauhiainen (T. Jauhiainen)
0000-0002-8227-5627 (H. Jauhiainen); 0000-0002-6474-3570 (T. Jauhiainen)

DHNB Publications, DHNB2023 Conference Proceedings, https://journals.uio.no/dhnbpub/issue/view/875

is made more difficult by the fact that there are damaged parts for which transliteration has sometimes been guessed, as well as grammatical additions by editors.

We have developed a highly accurate alignment method that uses a sequence alignment algorithm together with a dictionary of MdC - transliteration pairs generated initially from the intact words within the AES and the sentences in the RTC corpus that have the same number of words in both MdC and transliteration. Once the rest of the lines in the RTC were aligned, we could extract MdC - transliteration pairs and create transliteration models. Manual inspection of the models proved to be useful in helping to notice actual errors and inconsistencies in the RTC transliterations. We created additional rules to handle the errors and re-created the models. In this paper, we describe the automatic alignment pipeline and the openly published models.

Section 2 introduces some of the challenges in Digital Egyptology and gives further context on the research presented in this paper. In Section 3, we present previous work for the automatic alignment of hieroglyphs and some research using the same sequence alignment algorithm as this paper. Section 4 describes the AES - Ancient Egyptian Sentences and the Ramses Transliteration Corpora. We describe the alignment method itself in Section 5 as well as our evaluation of it using the Levenshtein distance [7]. In Section 6, we present some further steps that were needed to produce the model and the resulting models.

2. Background

Contrary to, for example, Assyriology, which has several online corpora of cuneiform texts,[1] Egyptology has no tradition of publishing machine-readable hieroglyphic texts [8]. In a hieroglyphic text, individual signs can be next to, above, or over another, and two or more signs can be nested. When Egyptologists study the texts, they draw a facsimile of the object and its inscriptions [9]. Many Ancient Egyptian texts were written by hand, which made the hieroglyphs more cursive, and Egyptologists generally transcribe these cursive signs into clearer hieroglyphs [10, 6]. Traditionally this was done by hand or by typesetting physical fonts, but nowadays, pictures produced with computers are also possible [6]. Applying optical character recognition (OCR) on hieroglyphic texts would produce machine-readable corpora. However, currently, the techniques are not readily usable as annotated texts in the same handwriting are needed for training the methods [8, 11]. When Egyptologists interpret hieroglyphic texts, they transliterate them with Latin letters and diacritics. The transliteration method used in Egyptology does not represent the text sign by sign, but instead, it is always an interpretation of the text as words.

It might seem obvious to use Unicode to produce machine-readable hieroglyphic texts. There are indeed over 1000 Unicode hieroglyphic characters, and since 2019, there are also so-called Format Control characters so that the signs can be presented properly in relation to each other [12, 13, 14].[2] However, the Unicode hieroglyphic characters are outside the Basic Multilingual

[1] E.g., Achemenet, http://www.achemenet.com; Open Richly Annotated Cuneiform Corpus (ORACC), http://oracc.museum.upenn.edu; ORACC at the Language Bank of Finland, http://urn.fi/urn:nbn:fi:lb-2019111601

[2] https://www.unicode.org/versions/Unicode14.0.0/ch11.pdf

Plane[3] and are not correctly handled by commonly used software applications. While the situation with the “digital divide” [16] between languages has improved as digital support for language diversity is expanding, many writing systems still have issues with common digital tools and technologies [17, 18]. Since the 1970s, egyptologists have been using special text editors to encode hieroglyphic texts so that the placement of the signs is maintained [13, 19]. Unfortunately, the machine-readable encodings have traditionally only been used for producing pictures of hieroglyphic texts and not published themselves [8].

Computer-assisted transliteration of hieroglyphic texts would aid Egyptologists in reading the encoded texts and publishing them for digital studies. Recently, Rosmorduc [20] proposed an automated transliteration method using neural networks. The method works on short sentences, but the sentence boundaries are not indicated in hieroglyphic texts, and sentences often span from one line to the next. A new method for transliterating whole texts at a time is, hence, needed. As Wiechetek et al. [21] have shown, using neural networks on less-resourced languages is often not feasible, and more traditional machine learning methods could be more effective. We have previously developed a back-off-based language identification method, HeLI [22], which still is superior to neural network-based identifiers in many tasks [23, 24]. Hence, the transliteration method we are currently developing is based on a similar back-off scheme, which at its core utilizes a transliteration model of hieroglyphic words and their transliterations together with the observed relative frequencies of the pairs.

3. Previous Work

3.1. Automatic Alignment of Hieroglyphs and Transliteration

The problem of automatic alignment of hieroglyphs and transliteration has been researched by Nederhof [25, 26]. His starting point for alignment differed from the current work since no word tokenization was present in the hieroglyphic texts he used. His method goes through the individual hieroglyphs and checks how well their possible transliterations match the existing transliteration. A customized scoring system gives varying penalties to different readings, and in the end, the reading with the lowest penalty is chosen.

In the corpus we use, the word tokenization is already indicated. Therefore it is possible to use much more straightforward methods for aligning the hieroglyphic words to their transliterated counterparts. By more straightforward, we mean that almost no expert knowledge in hieroglyphic writing is needed as our method learns the dictionaries automatically from the training corpus.

3.2. Sequence Alignment

In this work, we use the Needleman-Wunsch algorithm [27] to align encoded hieroglyphic text with their transliteration. This method was designed for and primarily used in bioinformatics. It works well, for example, when comparing proteins or genomic DNA sequences of up to tens of thousands of nucleotides [28]. The algorithm has also been used for aligning natural language text. Song et al. [29] evaluated several sequence alignment algorithms in identifying

[3] See the description by Tauber [15] for more information about the structure of the Unicode codespace.

sentence parallelism in student essays. They found the Needleman-Wunsch algorithm to perform best out of five individual algorithms but attained even better results when combining it with the others. In other areas of natural language processing, the Needleman-Wunsch algorithm has been used, for example, by Lai and Hockenmaier [30] and Itoh [31] to estimate semantic similarity in SemEval shared tasks. This method suits our needs exceptionally well as it supposes the sequences to be in the correct order but finds the places where items are missing from either sequence.

4. Corpora

In this section, we describe the two corpora from which we produced transliteration models AES - Ancient Egyptian Sentences [2], and The Ramses Transliteration Corpus [20].

4.1. AES - Ancient Egyptian Sentences

AES - Ancient Egyptian Sentences; Corpus of Ancient Egyptian sentences for corpus-linguistic research (AES) is a collection of more than 100,000 sentences with c. 254,000 annotated and transliterated words aligned with their MdC encoding [2]. The corpus was published in GitHub under a CC BY-SA license.[4] AES sentence corpus is formatted as JSON, which is more suitable for our pipelines than the TEI format of the Corpus of Egyptian Texts for the AED - Ancient Egyptian Dictionary (AED-TEI) [32],[5] from which AES has been extracted. AED-TEI contains more than 11,000 Egyptian texts and is itself based on a database snapshot from the "Strukturen und Transformationen des Wortschatzes der ägyptischen Sprache" project [1].[6] This database is also used to create the online version of the Thesaurus Linguae Aegyptiae (TLA) [33].[7]

Figure 1 shows an example of a JSON entry for a single Egyptian word in AES. A JSON entry consists of keyword – value pairs: *"keyword": "value"*. The AES information defines the value of the "mdc" keyword as an Egyptological transcription in MdC. What it actually means is that in the value of "mdc", the characters used in the transcriptions conform to the Buurman et al. 1988 transliteration scheme [4].[8] The MdC allows the use of these transliteration characters for some of the hieroglyphs in addition to the letter–number combinations which are the codes coming from the so-called *Gardiner Sign List* [34], the standard reference list for Egyptian hieroglyphs.[9] The MdC equivalent of the original hieroglyphic characters we are seeking is actually found in the value of the "hiero" keyword: M20-X1-Z2. We were hoping to use a dictionary created from the AES to align the MdC and the transliteration in the RTC corpus. However, the "mdc" value in AES is not similar to the RTC transliteration, as it includes extra annotation such as the plural marker ".pl" in the example. The value for the "written_form" keyword is described to be the same as for "mdc", but in Unicode, which seems to use the character set

[4] https://github.com/simondschweitzer/aes

[5] https://github.com/simondschweitzer/aed-tei

[6] https://nbn-resolving.org/urn:nbn:de:kobv:b4-opus4-29190

[7] https://thesaurus-linguae-aegyptiae.de

[8] For the currently most complete list of transliteration schemes for Ancient Egyptian, see https://en.wiktionary.org/wiki/Appendix:Egyptian_transliteration_schemes.

[9] https://www.unicode.org/notes/tn32/Unicode-MdC-Mapping-v1.utf8

```
{
    "_id": "IBUBdx3kjk1StkQar5VP5tBFvOk",
    "lineCount": "[x+7]",
    "written_form": "sḫ,t.pl",
    "mdc": "sx,t.pl",
    "cotext_translation": "Feld; Weide; Marschland",
    "lemma_form": "sḫ.t",
    "lemmaID": "141480",
    "zaehler": "14",
    "pos": "substantive",
    "numerus": "plural",
    "status": "st_absolutus",
    "hiero_inventar": "M20;X1;Z2",
    "hiero_unicode": "&#x131cf;&#x133cf;&#x133e5;",
    "hiero": "M20-X1-Z2"
},
```

Figure 1: Example of an Egyptian word in AES corpus.

of the Werning 2013 (traditional) transliteration scheme [35]. The value for the "lemma_form" keyword conforms to the Werning scheme and does not include extra annotation, and it is also possible to automatically transform the Werning transliteration in the "lemma_form" into Buurman-compliant transliteration. Thus, this is the entry we use in the automatic alignment experiments, even though it sometimes omits information such as the plural "w" in the case of the example in figure 1. The RTC equivalent transliteration for "M20-X1-Z2" would be "sx.wt".

The AES corpus is very heterogeneous. The dates attributed to the texts span from the Ancient Egyptian Old Kingdom to the Roman period, that is, for a period of over 2,500 years (c. 2600 BCE - 300 CE). Furthermore, the corpus is composed of texts from many different kinds of genres, such as religious texts, administrative texts, letters, medical texts, rock inscriptions, and so forth. Since the Egyptian language changed over time and different genres use different vocabulary, the model produced from the corpus is bound to be miscellaneous.

4.2. The Ramses Transliteration Corpus

The Ramses Transliteration Corpus, V. 2019-09-01 (RTC) was created by Rosmorduc [20] for training and testing his automated transliteration method [36]. The corpus was published in Gitlab[10] and Zenodo[11] under CC-BY-NC-SA license, and it contains sentences of encoded hieroglyphic text and respective lines of their transliterations in separate files [3]. Original hieroglyphic texts do not include word boundaries. However, since the data was originally collected in the Ramses Project [37] and is available for word searches in Ramses Online,[12] the corpus contains, in addition to texts without word boundaries, also separate versions of the files where words have been separated with underscores. In order to find word-transliteration pairs, we use the files with word boundaries. Examples of lines from these files can be seen in

[10] https://gitlab.cnam.fr/gitlab/rosmorse/ramses-trl

[11] https://doi.org/10.5281/zenodo.4954597

[12] http://ramses.ulg.ac.be

```
M17 Z7 _ M17 Z7 _ A1 _ D21 _ S29 G36 D21 N35A A2 _ M17 G17 _ I9 _
U23 G17 D21 G37 _ LACUNA V30 _ LACUNA
O4 Ff1 N35 G1 Ff1 Ff1 O1 F35 I9 D21 A1 _
G41 G1 _ O4 D21 Z7 N5 Z1 _ N35 _ MISSING G41 G1 _ W25 N35 W24 Z7 Y1 Z2 _
```

```
i w _ i w _ = i _ r _ s w r _ m _ = f _
m H r _ L A C U N A _ n b _ L A C U N A _
h y n - n f r _
p A _ h r w _ n _ m s _ p A _ i n w _
```

Figure 2: Four corresponding lines from the “src-sep-train.txt” and the “tgt-train.txt” files which are part of the RTC training set.

Figure 2. These examples also demonstrate the fact that what we refer to as a word actually means a token, such as “A1” corresponding to “=i”, which is a suffix pronoun. Both AES and RTC consider these as separate tokens from the words they are connected to.

Egyptian texts are often fragmentary and damaged in places. In RTC, there sometimes exists a possible transliteration for these damaged parts, whether individual signs or longer passages. Furthermore, grammatical forms not present in the hieroglyphic text have often been added in the transliteration. These guesses and additions have been marked in a variety of ways as transliterations have been produced by numerous scholars.

The RTC comprises over 71,000 sentences of Late Egyptian texts (c. 1550-1069 BCE). The texts are all from a single phase of the Ancient Egyptian Language, and the text types in Ramses Online are less varied than the ones in AES. The sentences are in random order, and there is no indication of which text they belong to. The training corpus contains 66,693 encoding and transliteration line pairs, whereas the validation set has 1,841 and the test set 2,729 line pairs.

All sets of the RTC were preprocessed to remove inconsistencies as far as possible. We inspected the sign values present in the training and the validation sets. As mentioned earlier, the Mdc allows using transliteration for certain signs, and, indeed, the corpus does include some annotations using transliteration rather than the Gardiner sign list keys for seven different signs.[13] In the first preprocessing step, we replaced them with the codes used elsewhere in the corpus for these hieroglyphs.[14]

Hieroglyphic texts in the RTC were originally written on papyri or so-called ostraca, that is, pieces of limestone or pottery. The texts that have survived until today have often suffered damage, particularly to their edges. When part of the text is missing or illegible, this is usually marked in the encoding. For missing parts of words, shading is often used. The Ramses corpus includes three different annotations, “SHADED1”, “SHADED2”, and “SHADED3”. There was no apparent reason for the different numbers, so we unified these three annotations as “SHADED”.

An entire illegible or missing word is in the corpus marked with “LACUNA” or “MISSING”. A “LACUNA” in the hieroglyphic text usually corresponds to a “L A C U N A” in transliteration. When a hieroglyph or word is marked as missing, it may correspond to transliteration or not. The encoding files do not indicate a token break after “LACUNA” or “MISSING” as is done in

[13] “n”, “A”, “f”, “t”, “i”, “nTrw”, and “nn”

[14] “N35”, “G1”, “I9”, “X1”, “M17”, “R8A”, and “M22 M22”

the transliteration, so we added an underscore after these in the encoding files used.[15]

5. Alignment Method

The Needleman-Wunsch algorithm [27] was originally developed to align long protein sequences. It finds the tokens that are the same, and when there is a mismatch, it either aligns the two tokens or inserts a gap finding the optimal alignment. To use the algorithm for hieroglyphic sentences, we need to align the transliterated words with the encoded words, and, therefore, we need a dictionary to compare these two. Our alignment method is, hence, based on creating a dictionary of pairs of encoded words and their transliteration and using it together with the Needleman-Wunsch algorithm. We tested three slightly different ways of producing the dictionary from words in the RTC assumed to be aligned without further processing. The dictionaries were tested with the various configurations of the Needleman-Wunsch algorithm on a manually aligned part of the validation set.

5.1. Gold Standard Alignment

In order to assess the efficacy of the developed method, we created a gold standard alignment from the RTC validation set. We created an alignment test set by choosing the first 100 pairs that had an unequal number of tokens. Dividing the tokens into columns in a Microsoft Excel spreadsheet provided an easy and fast way to manually verify and correct the existing alignments. An example of an encoding-transliteration pair in its original and manually aligned forms can be seen in Table 1.

orig.	D37 Z7	O29 O1 O1 Z1 G7	S34 U28 S29	D21	T10 X1 Z1 A1 Z3A	N35	...
	r d i	<w i >	p r - a A	a . w . s	r	<H r y >	...
gold	D37 Z7	-	O29 O1 O1 Z1 G7	S34 U28 S29	D21	-	...
	r d i	<w i >	p r - a A	a . w . s	r	<H r y >	...

Table 1
Example of original and correctly aligned encoding transliteration pair from the validation set.

5.2. Dictionaries

We experimented with using dictionaries created from both the RTC and the AES corpora with our implementation of the Needleman-Wunsch algorithm. The dictionaries are files consisting of tab-separated values for tokens together with all their possible transliterations. Figure 3 shows an entry for the token “M17 Z7” from the third RTC dictionary. The token has five possible transliterations “in-iw”, “ir”, “iw”, “iwf”, and “r”. The number in the third column indicates the frequency of the MdC – transliteration pair; however, this information is not used in the alignment algorithm.

[15] Examples of both “LACUNA” and “MISSING” can be seen in the second and fourth rows of the Figure 2.

```
M17 Z7  in-iw   1
M17 Z7  ir      3
M17 Z7  iw      3856
M17 Z7  iwf     1
M17 Z7  r       9
```

Figure 3: Example entry from the third RTC dictionary. The Egyptian token "M17 Z7" has five possible transliterations.

The dictionary from the AES corpus was built using each token that had a value for both the *hiero* and the *lemma_form* keywords. For our dictionary, the AES values written using the Werning scheme were automatically converted to the Buurman-compliant transliteration.

The three dictionaries from the RTC were built using the training set. After the preprocessing described in Section 4.2, quite a large portion of the encoding-transliteration pairs in the RTC training set had the same number of tokens in the encoding and the transliteration and were, therefore, assumed to be correctly aligned. This large number of correctly aligned pairs enabled us to create a token dictionary from the line pairs with an equal number of tokens with high confidence. When building the first and second versions of the RTC dictionary, we did not include any pairs of lines that had broken or partially broken transliteration, e.g., lines containing characters "[", "/", or "?", or inserted words that are not present in the encoding indicated by character "<" or parenthesis around the entire transliteration of a word.

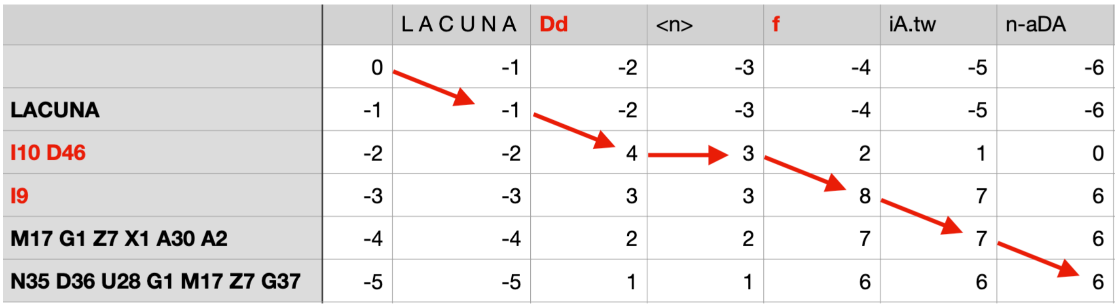
For the first RTC dictionary, we used pairs of lines that did not contain annotations "SHADED", "LACUNA", or "MISSING" in the encoding line. There were 31,938 such intact line pairs in the training material and the dictionary contained 32,023 words. To create the second version of the RTC dictionary, we did use the lines with "LACUNA" if none of the other indications of broken text or insertion were present. The total number of lines from which the second RTC dictionary, with 36,891 words, was created was 42,541. For the third RTC dictionary, we used all 56,752 pairs of lines with an equal number of tokens and gathered 45,225 words. However, when adding words to the third dictionary, we still left out the word pairs that contained the characters mentioned above or "SHADED" or "MISSING" as annotations on the encoding line.

5.3. Needleman-Wunsch

In the Needleman-Wunsch algorithm, a two-dimensional array represents all possible pairs of the units to be aligned. Only the pathways from the last items of the sequences to the first have to be considered. For each pair of encoded (x) and transliterated (y) sentences, we first construct an array of size $length\ of\ x + 1$ times $length\ of\ y + 1$ (see Fig. 4). The cells where $x = 0$ are filled with numbers descending from 0. The same is done to the cells where $y = 0$. The rest of the cells are filled according to Equation 1.

$$[x, y] = \max \begin{cases} [x-1, y-1] + \mathit{dscore} \\ ([x, y-1] - 1) \\ ([x-1, y] - 1) \end{cases} \tag{1}$$

That is, the cell is assigned the maximum of the scores in either the cell to the top left plus a dictionary score *dscore*, the cell to the left minus the penalty point (-1), or the cell above minus the penalty point (-1). The value of *dscore* is based on whether the transliteration is found in the dictionary entry for the encoding of the words under consideration (5 points) or not (-1 points). In Figure 4, only two encoded words (I10 D46 and I9) have a matching transliteration (Dd and f, respectively) in our dictionary. For all the other pairs, the *dscore* is -1.



		L A C U N A	Dd	<n>	f	iA.tw	n-aDA
	0	-1	-2	-3	-4	-5	-6
LACUNA	-1	-1	-2	-3	-4	-5	-6
I10 D46	-2	-2	4	3	2	1	0
I9	-3	-3	3	3	8	7	6
M17 G1 Z7 X1 A30 A2	-4	-4	2	2	7	7	6
N35 D36 U28 G1 M17 Z7 G37	-5	-5	1	1	6	6	6

Figure 4: The Needleman-Wusch algorithm used on a encoding - transliteration sentence pair. The optimal alignment is found by backtracing the arrows that indicate from where the value assigned to the cell was received. Only the word pairs marked with red are found in our dictionary and get the *dscore* 5. See Table 2 for the final alignment.

Another array is constructed simultaneously and filled with the information from which direction – left, top, or top left – the maximum score was found. In Figure 4, the relevant directions are conveyed with arrows. After both arrays have been filled, the best alignment is found by starting from the cell [*length of x*][*length of y*], that is, from the bottom right cell. The encodings and transliterations are aligned, starting from the last word and moving toward the first. If the score came from the top left, the encoding and transliteration x and y are aligned. If the score was received from the left, the transliteration is aligned with an empty slot, marked by "-". When the score comes from the above cell, it is the transliteration that is left empty. One then moves on to the cell where the maximum score came from until reaching the top left corner of the array [$x = 0$][$y = 0$]. For the alignment of the sentences in Figure 4 see Table 2.

Encoding	LACUNA	I10 D46	-	I9	M17 G1 Z7 X1 A30 A2	N35 D36 U28 G1 M17 Z7 G37
Transliteration	L A C U N A	D d	<n>	f	i A . t w	n - a D A

Table 2
The final alignment of the sentence pair in Fig. 4. In the figure, the word pair I10 D46 - <n> gets its score from the cell to its left (indicated by the horizontal arrow) and, therefore, the transliteration <n> has been aligned with an empty slot.

When the sentence is long or has many words not present in the dictionary, the maximum score for a cell may sometimes be received from several directions; that is, there are multiple optimal alignments. In order to find out which passage is best, we sum up the *dscores* for each pair on every possible passage. The highest scored alignment is then returned. In case there are several alignments with the highest score, we favor the diagonal alignment and, since additional words are more common in transliteration, left over top.

5.4. Evaluation of the Segmentation Method

We evaluated the method using the Levenshtein distance, which calculates the minimum number of single token edits to change one string to the other.[16] Levenshtein distance was also used for assessing the correctness of transliteration of hieroglyphic texts by Rosmorduc [20]. We implemented the method according to "Algorithm X" described by Wagner and Fischer [38]. Like in the Needleman-Wunsch algorithm, when comparing strings x and y, one uses a two-dimensional array, but the cells where $x = 0$ are filled with numbers ascending from 0 instead of descending. The same is done to the cells where $y = 0$. Like in the Needleman-Wusch algorithm, when assigning a score to a cell, one always considers the score in the cells to the left, top left, and above. A penalty point is added to the scores to the left and above and, if the tokens to compare are not the same, also to the score in the top left cell. After filling the array, the Levenshtein distance is retrieved from the cell [*length of x*][*length of y*]. We used one as a penalty point. We compared both lines produced with our alignment method, that is the encoding line and the transliteration line to the respective ones in the gold standards. This means that, in fact we always compared 200 lines. The smaller the Levenshtein distance, the more similar the lines are.

The Levenshtein distance of our alignment test set to our gold standard was 124 before doing any aligning. Examining the lines showed there were sometimes long sections of insertions, even at the beginning and the end of the sentences.

Initially, in our Needleman-Wunsch algorithm, we simply checked whether the transliteration was in the entry of the encoded word we were trying to align it with. Since a hieroglyphic word could be written in many ways by using different sign combinations, we tested various slightly different ways of finding out whether the encoding and the transliteration refer to the same word.

The basic matching scheme in the alignment algorithm was to compare the transliteration to be aligned with all possible transliterations for the encoded hieroglyphic token. In case of a match, the *dscore* is 5. Already in the initial matching scheme, the "Simple", if there was no match for the first rule, the alignment was additionally considered a match if the transliteration or any of the possible transliterations started with and completely contained the other. This partial match gave a *dscore* of 4. Using the simple scheme, the performance of the AES dictionary was worse than expected, giving a distance of 147, which means there were more mistakes than in the original alignment. Using our first Ramses dictionary gave a distance of 40, which was topped by 30 by the second dictionary and 26 by the third one (see Table 3). Adding AES to each of the Ramses dictionaries improved the results slightly to 38, 28, and 24, respectively.

In our second matching scheme, we added the "First 3" rule, i.e., taking the first three MdC codes from the encoding under scrutiny and adding the transliterations of all the words starting with those three codes to the list of possible ones before checking. If neither of the simple scheme rules matched, the "First 3" rule gave a *dscore* of 4. The distance with the AES dictionary improved to 98 while the first Ramses dictionary gave 16, and the second and the third 10. This rule seemed to work especially while using the AES dictionary with the first and second Ramses

[16] See example analyses by Kruskal [7] (Section 4. Levenshtein And Other Distances) on how to calculate the Levenshtein distance.

Dict.	Distance Simple	+First 3	+ Logogram
AES	147	98	81
RTC 1st	40	16	8
RTC1 + AES	38	12	6
RTC 2nd	30	10	2
RTC2 + AES	28	6	2
RTC 3rd	26	10	2
RTC3 + AES	24	6	2

Table 3
Evaluation of the the first 100 lines of the validation set with unequal number of tokens against our gold standard using the first and second dictionaries with various matching methods in the Needleman-Wunsch algorithm.

dictionaries. The first RTC gave 12, and the second 6. The rule had no effect with the third dictionary compared to the second.

In the third matching scheme, a "Logogram" rule was added. With this rule, only the first sign of the encoding is considered. We used the sign similarly to the second rule of the "Simple" scheme. If none of the previous rules was usable, the "Logogram" rule gave a *dscore* of 3. This resulted in a small improvement of the Levenshtein distance with all the dictionaries: AES 81, the first Ramses dictionary 8, and the combination of these 6. All the other dictionaries gave the distance 2.

6. Transliteration Models

Since we wanted to align the sentences as well as possible and produce good-quality transliteration models, but we could not achieve Levenshtein distance 0 in our evaluation, we resorted to using a corpus-specific addition in the alignment method. The *dscore* was set to -5 if the transliteration contained any of the characters indicating an insertion by the annotator, i.e., parenthesis or angle brackets around the word. These could be possibly preceded by "=", indicating that a suffix pronoun was meant. This measure gave the distance 0 with all three Ramses dictionaries, even with the simple matching scheme.

In order to verify the reliability of the alignment of sentence pairs with an uneven number of words in the RTC, we used the third Ramses dictionary and all the matching rules, including the corpus-specific addition, for aligning the entire Ramses training set. We then made a wordlist out of the aligned lines and studied that manually. The broken words and additions are omitted from the word lists used for producing the transliteration model, but they were retained for this step of the process. When intact words were found to be paired with a '-', we studied the sentence where that happened. We found several sentences where the MdC and transliteration did not match. For example, some names of kings were always written as one word in the encoding but as two words in the transliteration. We made rules to align these names properly. For creating the transliteration model, we added these rules to the preprocessing step. Occasionally, some of the hieroglyphs had also been left untransliterated, and sometimes there were transliterations that had been deduced from the context but with no counterpart in the MdC. Often this did not matter as our method could detect the missing words, but for some sentences, we just could not align properly, so we made a list of sentences to ignore.

Once we were content with the dictionary built from the aligned lines, we used the dictionary to build the transliteration model. The various dictionaries contain information on the frequency of each MdC-transliteration pair and can be used as transliteration models of the specific corpora they were built from. In order to publish the models in a more structured format, we made a script to write them using the JSON scheme. The script allows one to build a transliteration model from a desired number of word lists. We build JSON-format models from the AES words, the words in the Ramses training model, and a combined AES-Ramses model.

```
{
      "encoding": "M17 Z7",
      "interpretations": [
        {
          "transliteration": "iw",
          "freq": 8142,
          "relFreq": 99.75
        },
        {
          "transliteration": "r",
          "freq": 13,
          "relFreq": 0.16
        },
        {
          "transliteration": "ir",
          "freq": 4,
          "relFreq": 0.05
        },
        {
          "transliteration": "in-iw",
          "freq": 2,
          "relFreq": 0.02
        },
        {
          "transliteration": "iwf",
          "freq": 1,
          "relFreq": 0.01
        }
      ]
    },
```

Figure 5: Example of an Egyptian word in the transliteration model created from the RTC training corpus.

The models have been published under an open license on Zenodo and GitHub.[17] The models are JSON files with separate entries for each MdC token, as seen in Figure 5. The token in the example, "M17 Z7", is the same as in Figure 3. The "encoding" keyword gives the MdC for the

[17] http://doi.org/10.5281/zenodo.7991240, https://github.com/MaReTEgyptologists/TranslitModels

token, and all attested transliterations are collected under the "interpretations" keyword. For each transliteration, the frequency "freq" is indicated as is the observed probability, "relFreq", of the transliteration for the given MdC token. Due to being able to align the different length MdC – transliteration pairs, the number of "iw" transliterations for "M17 Z7" more than doubled even though the number of lines only grew from c. 57,000 to 67,000. This difference is due to the fact that misalignment occurs more often in longer sentences than in short ones.

The Table 4 shows the sizes of the different versions of the published models.

Model	# of unique tokens	Total frequency	# of unique MdC – transliteration pairs
AES	43,416	250,909	46,811
RTC	48,914	447,719	55,049
AES + RTC	84,558	698,628	97,777

Table 4
The published transliteration models and their sizes.

7. Conclusions

Our alignment method manages to align even complex sentences with the difference of several words in the original encoding and transliteration lines. Increasing the size of the dictionary by including more pairs of words improved the results every time. Using a penalty for the obvious insertions by annotators was needed to align the original lines and our gold standard perfectly. In order to build good-quality transliteration models, some additional corpora-specific rules were needed. Although the alignment method depends on the corpus format, which is currently unique to this export done by Rosmorduc [20], we have published the software used for building the dictionaries and aligning the sentences. They cannot be used off-the-shelf for other corpora, but we believe that they can be modified for or at least give an example of how the Needleman-Wunsch can be used for aligning sentences that do not have the same format.

We have published a combined transliteration model by using the words from the AES and the RTC corpora. However, our attempts to use the AES for aligning MdC and transliteration sentence pairs in the RTC corpus were unsuccessful and showed that the corpora are very different. We assume that the separate models built from each of the corpora will be more useful.

The most obvious thing for future work would be to evaluate the alignment method using the test set in the Ramses Transliteration Corpus. However, as we wish to use the test set to evaluate the forthcoming automatic transliteration method, we are hesitant to look at it, which is needed for manual alignment.

We have used the RTC Transliteration Model in our first experiments on automatically segmenting hieroglyphic texts into words, a task that is needed before automated transliteration. We noticed that information on two or more words in a window might be useful for the task. Hence, we intend to produce additional token n-gram transliteration models from the texts we have aligned.

Acknowledgments

The Kone Foundation has funded this research. This work has been partially supported also by the Academy of Finland (Funding decision no. 341798). We thank Serge Rosmorduc for publishing the Ramses Transliteration Corpus and Simon Schweitzer for publishing the AES corpus with an open license. We are also grateful to the people behind the Ramses Online and Thesaurus Linguae Aegyptiae services for allowing the data produced in their projects to be used in research.

References

[1] T. S. Richter, I. Hafemann, H.-W. Fischer-Elfert, P. Dils, Teilauszug der Datenbank des Vorhabens ”Strukturen und Transformationen des Wortschatzes der ägyptischen Sprache” vom Januar 2018, 2018. URL: https://nbn-resolving.org/urn:nbn:de:kobv:b4-opus4-29190.

[2] Schweitzer, Simon D., AES - Ancient Egyptian Sentences; Corpus of Ancient Egyptian sentences for corpus-linguistic research, 2021. URL: https://github.com/simondschweitzer/aes.

[3] University of Liege/Project Ramses, The Ramses transliteration corpus V. 2019-09-01, 2020. URL: https://gitlab.cnam.fr/gitlab/rosmorse/ramses-trl.

[4] J. Buurman, N. Grimal, M. Hainsworth, J. Hallof, D. van der Plas, Inventaire des signes hiéroglyphiques en vues de leur saisie informatique: Manuel de codage des textes hiéroglyphiques en vue de leur saisie sur ordinateur, Institut de France, Paris, 1988.

[5] M.-J. Nederhof, A Revised Encoding Scheme for Hieroglyphic, in: Proceedings of the XIV Computer-aided Egyptology Round Table, Pisa, Italy, July 2002, IE2002, 2002.

[6] S. Polis, S. Rosmorduc, Réviser le codage de l'égyptien ancien: vers un répertoire partagé des signes hiéroglyphiques, Document numérique 16 (2013) 45–67.

[7] J. B. Kruskal, An overview of sequence comparison: Time warps, string edits, and macromolecules, SIAM Review 25 (1983) 201–237. URL: https://epubs.siam.org/doi/pdf/10.1137/1025045.

[8] M.-J. Nederhof, OCR of Handwritten Transcriptions of Ancient Egyptian Hieroglyphic Text, in: Proceedings of Altertumswissenschaften in a Digital Age: Egyptology, Papyrology and Beyond, 2015.

[9] C. Thiers, The So-Called Karnak Method, in: The Oxford Handbook of Egyptian Epigraphy and Paleography, Oxford University Press, 2020. doi:10.1093/oxfordhb/9780190604653.013.21.

[10] A. H. Gardiner, The transcription of new kingdom hieratic, The Journal of Egyptian Archaeology 15 (1929) 48–55. URL: http://www.jstor.org/stable/3854012.

[11] B. Bermeitinger, S. A. Gulde, T. Konrad, How to compute a shape: Optical character recognition for hieratic, in: C. Gracia Zamacona, J. Ortiz-García (Eds.), Handbook of Digital Egyptology: Texts, Editorial Universidad de Alcalá, 2021, pp. 121–138.

[12] The Unicode Consortium, The Unicode Standard, Version 14.0.0, 2021. URL: https://www.unicode.org/versions/Unicode14.0.0/.

[13] R. B. Gozzoli, Hieroglyphic Text Processors, Manuel de Codage, Unicode, and Lexicogra-

phy, in: S. Polis, J. Winand (Eds.), Texts, Languages Information Technology in Egyptology, Presses Universitaires de Liège, Liège, 2013, pp. 89–101.

[14] M.-J. Nederhof, S. Polis, S. Rosmorduc, Unicode control characters for ancient egyptian, in: Proceedings of the International Congress of Egyptologists XII, F.R.S.-FNRS - Fonds de la Recherche Scientifique, IFAO, In press.

[15] J. K. Tauber, Character encoding of classical languages, in: M. Berti (Ed.), Digital Classical Philology, De Gruyter Saur, Berlin, Boston, 2019, pp. 137–158. doi:doi:10.1515/9783110599572-009.

[16] R. Cullen, Addressing the digital divide, Online information review 25 (2001) 311–320.

[17] F. Albarillo, Evaluating language functionality in library databases, International Information & Library Review 48 (2016) 1–10. doi:10.1080/10572317.2016.1146036.

[18] I. A. Zaugg, Digital inequality and language diversity: An ethiopic case study, in: M. Ragnedda, A. Gladkova (Eds.), Digital Inequalities in the Global South, Springer International Publishing, 2020, pp. 247–267. doi:10.1007/978-3-030-32706-4_12.

[19] S. Rosmorduc, Digital writing of hieroglyphic texts, in: C. Gracia Zamacona, J. Ortiz-García (Eds.), Handbook of Digital Egyptology: Texts, Editorial Universidad de Alcalá, 2021, pp. 37–53.

[20] S. Rosmorduc, Automated Transliteration of Late Egyptian Using Neural Networks: An Experiment in "Deep Learning", Lingua Aegyptia-Journal of Egyptian Language Studies 28 (2020) 233–257.

[21] L. Wiechetek, K. Hiovain-Asikainen, I. L. S. Mikkelsen, S. Moshagen, F. Pirinen, T. Trosterud, B. Gaup, Unmasking the myth of effortless big data - making an open source multi-lingual infrastructure and building language resources from scratch, in: Proceedings of the Language Resources and Evaluation Conference, European Language Resources Association, Marseille, France, 2022, pp. 1167–1177. URL: https://aclanthology.org/2022.lrec-1.125.

[22] T. Jauhiainen, K. Lindén, H. Jauhiainen, HeLI, a word-based backoff method for language identification, in: Proceedings of the Third Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial3), The COLING 2016 Organizing Committee, Osaka, Japan, 2016, pp. 153–162. URL: https://www.aclweb.org/anthology/W16-4820.

[23] T. Jauhiainen, H. Jauhiainen, K. Lindén, HeLI-OTS, off-the-shelf language identifier for text, in: Proceedings of the Thirteenth Language Resources and Evaluation Conference, European Language Resources Association, Marseille, France, 2022, pp. 3912–3922. URL: https://aclanthology.org/2022.lrec-1.416.

[24] B. R. Chakravarthi, M. Gaman, R. T. Ionescu, H. Jauhiainen, T. Jauhiainen, K. Lindén, N. Ljubešić, N. Partanen, R. Priyadharshini, C. Purschke, E. Rajagopal, Y. Scherrer, M. Zampieri, Findings of the VarDial evaluation campaign 2021, in: Proceedings of the Eighth Workshop on NLP for Similar Languages, Varieties and Dialects, Association for Computational Linguistics, Kyiv, Ukraine, 2021, pp. 1–11. URL: https://www.aclweb.org/anthology/2021.vardial-1.1.

[25] M.-J. Nederhof, Automatic Alignment Of Hieroglyphs And Transliteration, Gorgias Press, 2009, pp. 71–92. doi:doi:10.31826/9781463216269-007.

[26] M.-J. Nederhof, Automatic Creation of Interlinear Text for Philological Purposes, Traitement Automatique des Langues (2009).

[27] S. B. Needleman, C. D. Wunsch, A General Method Applicable to the Search for Similarities in the Amino Acid Sequence of Two Proteins, Journal of Molecular Biology 48 (1970) 443–453. doi:https://doi.org/10.1016/0022-2836(70)90057-4.

[28] A. L. Delcher, S. Kasif, R. D. Fleischmann, J. Peterson, O. White, S. L. Salzberg, Alignment of Whole Genomes, Nucleic Acids Research 27 (1999) 2369–2376. doi:10.1093/nar/27.11.2369.

[29] W. Song, T. Liu, R. Fu, L. Liu, H. Wang, T. Liu, Learning to Identify Sentence Parallelism in Student Essays, in: Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers, The COLING 2016 Organizing Committee, Osaka, Japan, 2016, pp. 794–803. URL: https://aclanthology.org/C16-1076.

[30] A. Lai, J. Hockenmaier, Illinois-LH: A Denotational and Distributional Approach to Semantics, in: Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014), Association for Computational Linguistics, Dublin, Ireland, 2014, pp. 329–334. URL: https://aclanthology.org/S14-2055. doi:10.3115/v1/S14-2055.

[31] H. Itoh, RICOH at SemEval-2016 Task 1: IR-based Semantic Textual Similarity Estimation, in: Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016), 2016, pp. 691–695.

[32] Schweitzer, Simon D., Corpus of Egyptian Texts for the AED - Ancient Egyptian Dictionary, 2019. URL: https://github.com/simondschweitzer/aed-tei.

[33] S. D. Schweitzer, Compiling the lexicon of ancient egyptian: State of the art, in: C. Gracia Zamacona, J. Ortiz-García (Eds.), Handbook of Digital Egyptology: Texts, Editorial Universidad de Alcalá, 2021, pp. 103–120.

[34] Sir A. Gardiner, Egyptian Grammar: Being an Introduction to the Study of Hieroglyphs, 3rd. ed., Griffith Institute, Oxford, 1957.

[35] D. A. Werning, Einführung in die hieroglyphisch-ägyptische Schrift und Sprache: Propädeutikum mit Zeichen-und Vokabellektionen, Übungen und Übungshinweisen, Humboldt-Universität zu Berlin, 2015.

[36] S. Rosmorduc, Ramses automated translitteration software, 2021. doi:10.5281/zenodo.4954597.

[37] J. Winand, S. Polis, S. Rosmorduc, Ramses. An Annotated Corpus of Late Egyptian, in: Proceedings of the 10th International Congress of Egyptologists. University of the aegean, rhodes 22-29 May 2008, Peeters, 2015, pp. 1513–1521.

[38] R. A. Wagner, M. J. Fischer, The String-to-String Correction Problem, J. ACM 21 (1974) 168–173. doi:10.1145/321796.321811.

A Sustainable West? Analyzing Clusters of Public Opinion in Sustainability Western Discourses in a Collection of Multilingual Newspapers (1999-2018)

Elena Fernández Fernández[1], Germans Savcisens[2]

[1]*Institute of Computational Linguistics (Univesity of Zurich), Andreas Strasse 15, Zurich, 8050, Switzerland*
[2]*Section for Cognitive Systems (Technical University of Denmark), Building 324, Kongens Lyngby 2800, Denmark*

Abstract
In this article, we analyze the temporal and geographic evolution of sustainability-related discourses over a time frame of twenty years (1999-2018). We use a collection of multilingual newspapers in English, French, German, Spanish, and Italian, as a proxy. We filter documents using four key terms: sustainable development, climate change, environment, and pollution, seeking to explore how different newspapers encode the same message, aiming to detect points of contact (agreement) and rupture (polarity). Our methodology includes Topic Modelling (Pachinko Allocation [1]), word embeddings [2], Ward's hierarchical cluster analysis [3], and network analysis [4]. Our results show a progressive simplification of semantic fields over time, reflecting less polarizing views across countries and, therefore, showing an increasing agreement on sustainability-related discourses in our contemporary societies. Moreover, we also notice little variation of newspapers rhetorics over time. Therefore, this article also contributes with a meta-reflection about newspapers behaviour as information containers.

1. Introduction

The United Nations 2030 Agenda for Sustainable Development (United Nations [5]), organized into seventeen sustainable development goals, is a clear indicator that signals contemporary concerns about the necessity of taking various measures in the present to ensure a well-balanced growth of society. While relatively new, discourses about sustainability have received critical attention across different fields (Drucker [6]). However, we believe that the analysis of the historical and geographical evolution of sustainability-related concepts has not been accomplished yet with enough granularity.

In this article, we address that research gap by using a multilingual dataset of contemporary newspapers in English (*The Times, The New York Times International, Chicago Daily Herald, The Irish Times*), German (*NZZ*), French (*Le Figaro*), Italian (*La Stampa*), and Spanish (*El País*) over a period of twenty years (1999-2018). Our goal is to analyze the temporal evolution of geographic clusters of public opinion in Western societies seeking to detect points of convergence and

DHNB2023: Digital Humanities in the Nordic and Baltic Countries 2023. Sustainability, Environment, Community, Data. March 08-10, 2023, Online Conference

elena.fernandezfernandez@uzh.ch (E. F. Fernández); gersa@dtu.dk (G. Savcisens)
https://elenafernandezfern.wixsite.com/elena-fernandez (E. F. Fernández);
https://carlomarxdk.github.io/portfolio/ (G. Savcisens)
0000-0001-6596-6349 (E. F. Fernández); 0000-0002-5811-3230 (G. Savcisens)

DHNB Publications, DHNB2023 Conference Proceedings, https://journals.uio.no/dhnbpub/issue/view/875

rupture. To do so, we first filter our dataset by selecting documents that contain three key terms aligning with three of the seventeen United Nations Sustainable Development Goals. Thus, we firstly filter documents using the key terms climate change (which is related to Goal Thirteen, Climate Action), pollution (which is similar to Goal Six, Clean Water, and Sanitation, and to Goal Twelve, Responsible Consumption and Production), and environment (related to Goal Fifteen, Life on Land). We have selected these terms as an initial exploratory approach to analyzing press coverage on sustainability discourses. On top of those three key terms that are semantically related to three of the United Nations Sustainable Development Goals, we also select the concept of sustainable development to inspect the discussion about this topic in a broader context.

To uncover the temporal variations in data, we split the dataset into epochs of five years (1999-2003, 2004-2008, 2009-2013, and 2014-2018). We implement a three-step methodology consisting of Topic Discovery, Topic Evolution Analysis, and Sentiment Analysis. The first two aim to capture the contextual patterns across time and space, while the last one aims to identify the sentimental direction of the discourse.

This article contributes to establishing intellectual bridges between the emerging field of Environmental Humanities, big data and computational methods, and current efforts to create a sustainable world. Environmental Humanities acts as an umbrella term that gathers a variety of interdisciplinary research lines aiming to demystify beliefs that nature crises are a one-sided consequence of technology, therefore ignoring the role of history and culture as actors in that process (Heise et al. [7]): "The environmental humanities, by contrast, envision ecological crises fundamentally as questions of socioeconomic inequality, cultural difference, and divergent stories, values, and ethical frameworks". (2). In the following pages, we will analyze quantitatively public discourses about sustainability in Western societies under that theoretical framework using newspapers as a proxy.

While our work is informed by both Habermas [8] (who considers newspapers as the embodiment of the public sphere where private citizens evaluate government actions and make informed decisions about who to vote in democratic elections) and Herman and Chomsky [9] (who, on the other hand, propose the so-called "propaganda model," that states that during the last decades, private corporations and governments have progressively taken control over mass media worldwide), we do not wish to participate in that debate. However, we are very interested in observing how citizens living in diverse Western societies are exposed to information similarly or differently over time. We believe that an empirical analysis of cultured discourses about sustainability may be beneficial to gain a better understanding of the social effect of the press, laying the ground for further analyses across sectors. In doing so, we dialogue with Hay et al. [10], who state that knowledge is a driver of human action towards sustainability. In this article, we explore the temporal and geographic evolution of those discourses in Western societies to assess the footprint of both culture and zeitgeist in environmental communication. Moreover, we reflect about newspapers as containers of information, arguing that big data and computational methods have the capacity to shed light in a yet unseen manner about the segment of reality they represent.

2. Related Work

State of the art across disciplines has analyzed qualitatively and quantitatively discourses about sustainability under a variety of theoretical perspectives using different corpora. Qualitatively, (Philippon [11]) analyzes sustainable food rhetorics (i.e., the slow food movement), (Lenz [12]) explores points of convergence and divergence in discourses about digital technologies and the future of sustainability internationally and across sectors, and (Beling et al. [13]) discusses tensions in Ecological Economics sustainable development discourses between the Euro-Atlantic cultural region and the Global South north. Moreover, in the field of sustainability communication (defined by (Godemann and Michelsen [14]) as the development of critical awareness about the relationship between humans and their environment in social discourse), Brand [15] observes different frames of sustainability discourses from a sociological perspective in Germany, (Kruse [16]) analyzes human behaviour towards sustainability under the frame of environmental psychology, and (Witt [17]) explores Media Theory and Sustainability Communication. We believe that our work complements these approaches by focusing on the discursive aspects of sustainability (yet using distant reading methodologies). Environmental Humanities is a relatively recent area of scientific interest, and, to our knowledge, there are not many studies focusing on the historical evolution of sustainability rhetorics. In this paper, we seek to fill that research gap.

Quantitatively, Barkemeyer et al. [18] use 115 leading national newspapers from 39 different countries to inspect public opinion of sustainability concepts in the field of finance - their data covers publications over the eighteen-year period, i.e., 1990-2008. They filter the corpus using key terms such as sustainability, sustainable development, and business ethics, as well as a variety of tokens in the semantic family of business responsibility. Their methodology quantifies the number of appearances of key terms historically as a proxy for collective attention of environmental-related discourses, noting a clear increment over time. They continue this line of work in Barkemeyer et al. [19], this time inspecting different media coverage between the Global North and South of sustainability-related terms. Their corpus consists of articles from 115 multilingual newspapers based in 41 countries in eight different languages using the year 2008 as an observational time. They find significant differences in coverage of the selected tokens in both regions, reflecting diverging public policy sustainability agendas in developed and developing countries. Similarly, Kumar and Das [20] analyze sustainability reports of 200 firms internationally between 2008 and 2017. They quantify the number of appearances of 208 key terms related to sustainability in their corpus, noting a general trend of an increasing presence over countries across time. In the area of Media and Communication studies, Fischer et al. [21] analyze the semantic evolution of the terms sustainable and sustainability in German (choosing the key term *nachhaltig* in six German newspapers between 1995 and 2015. Results show an increasing number of mentions of the key terms across newspapers, as well as a progressive semantic technification of the sustainability field as articles approached contemporary times. We are interested in dialoguing with these lines of work, and we further investigate these findings by extending their observational time (we will use two decades, 1999-2018) as well as by specifically inspecting points of contact (similar encoding across multilingual newspapers of a same environmental-related concept) and rupture (different expression of a same message) in our dataset.

In the areas of Environmental Management and Public Policy, Sebestyén et al. [22] use the Voluntary National Reviews (mandatory reports that countries worldwide must accomplished after the declaration of the UN Sustainable Development Goals) of 75 countries in English published between 2016 and 2021. Implementing network analysis, they discover clear geographic clusters of data, identifying a Southern European cluster composed of Italy, Spain, Montenegro, Greece, and Portugal; a Northern European one formed by France, Germany, and Sweden; and two less clearly geographically connected ones, yet still in close proximity (a first one containing Samoa, Egypt, and Azerbaijan, while a second one including The Maldives, Georgia, India, Belize, Turkey, and Tajikistan). In a similar line of research (and of particular interest for this paper) is the work of Hallin and Mancini [23], where the authors propose the existence of geographic clusters of media models rooted in different historical, economic, and political traditions. They identify three main regions of media behaviour. Firstly, they point out to a Mediterranean knot, composed of Spain, Greece, and Portugal (sharing very similar contemporary history patterns of military dictatorships during the twentieth century and, therefore, a late-blooming of democracy), Italy, and to a lesser degree France. Secondly, they identify a North-Central European region composed of Scandinavia, the low countries, Germany, Austria, and Switzerland, claiming a similar shared recent history as well as many cultural, economic, and political traits. Thirdly, they identify a third media area, the North-Atlantic Anglo-speaking region, formed by the UK, Ireland, the United States, and Canada, following a similar line of reasoning.

Big data and computational methods, under the analytical frame of environmental humanities, can enrich these academic discussions by providing new findings to inspect those two research dimensions (multilingualism and historical perspective). In this paper, we seek to identify a possible existence of geographic clusters of media discourses in Western societies as well as their plausible temporal evolution during the last twenty years. We believe that doing so will provide innovative perspectives about the dialectical relationship between public opinion and sustainability discourses, shedding light on their reversed influence across time and space. Moreover, we contribute with cutting-edge discoveries to current discussions that investigate the historic multicultural evolution of the press in Western societies. Indeed, our main contribution consists of a novel observation of the textual structure of sustainability-related discourses in newspapers over the last twenty years. We are able to state whether sustainability rhetorics have become more fragmented (showing geographic polarization) or more homogeneous (showing an increasing international agreement) over time, as well as observing patterns of information diversity and emotions in the press of our selected countries.

3. Dataset

Our multilingual dataset is composed of six different newspapers of record (*Le Figaro, El País, La Stampa, NZZ, The Times, Irish Times*, one international newspaper (*The New York Times International*), and one local newspaper (*Chicago Daily Herald*). We believe that it is interesting to compare and contrast discourse dynamics across those three sorts of news outlets. Newspapers of record are considered relatively neutral vehicles of public opinion (i.e., leaning towards center political views) and, therefore, understood as non-partisan units of recorded history. *The New*

Table 1
Number of articles per newspaper (for the period 1999-2018)

The Times	The NYT	Chicago	Irish Times	Le Figaro	El Pais	La Stampa	NZZ
48985	26181	23123	43341	32545	48819	37616	23990

York Times International (previously known as *The International Herald Tribune*), is a newspaper that, from its inception, was conceived as an international news outlet focusing on international, economic, and cultural news (Sterling [24], Greenslade [25]). *Chicago Daily Herald* is a local newspaper yet headquartered in a major city with an international economic weight. We think that, for knowledge discovery purposes, it would be relevant to observe points of contact and rupture in sustainability-related discourses not only geographically and temporally but also across different news outlets publications, and editorial policies.

Our corpus has a total of 261477 documents, consisting of 118507 in English, 32545 in French, 48819 in Spanish, 37616 in Italian, and 23990 in German. We use the database Datalab, powered by LexisNexis, to retrieve the data. Each article includes a title, date of publication, as well as the full text (and other accompanying metadata). Table 1 shows the number of documents per newspaper.

Datalab provides access to a variety of newspapers internationally in more than twenty languages with a time coverage of approximately thirty years (1990-2020, although each newspaper presents a different time availability). Datalab is a news repository linked to a LexisNexis owned Jupyter Notebooks environment where users can execute data analyses using either Python or R programming languages. Datalab provides access to all the articles whose publication rights are owned by the newspapers they host. But, articles written by freelancers are not owned by the newspaper, and thus, they do not appear in the database. Consequently, the number of yearly articles available in the database may not necessarily coincide with the real number of yearly articles published by newspapers. While we are aware that this may represent a bias in our data analysis, we still believe that we can observe general trends in the geographic and temporal evolution of sustainability discourses.

We filter each newspaper using four key terms: Sustainable Development, Climate Change, Pollution, and Environment. That means that a) we create four different datasets per newspaper (each dataset matches one of the four key terms), and b) that we only select documents that contain those key terms. We use each word's native language root form (i.e., stem) to filter the documents (Porter [26]). Table 2 shows each word with its corresponding root. We have followed two different criteria to translate each term into different languages. Firstly, we have consulted how multilingual versions of Wikipedia articles translate our selected key terms in different languages. For example, to infer how climate change is translated into French, we consult the English Wikipedia page of climate change and then select the French version of the same article seeking to observe how Wikipedia users express the same concept in French. Oftentimes, we encounter that there are several synonyms of the same concept. For instance, we note that there are three options to translate climate change into French: *réchauffement climatique, changement climatique, dérèglement climatique*. Aiming to decide which one of them is most semantically relevant in press discourses, we filter our corpus with each one in order to observe the one that produces the higher number of documents. In the case of French,

Table 2
Table of filtering queries for each language and key term

	Queries			
Terms	***Sustainable Development***	***Pollution***	***Climate Change***	***Environment***
EN	sustain! AND develop!	pollut!	climat! AND chang!	environ! AND natur!
FR	développ! AND souten!	pollut!	chang! AND climat!	environ! AND naturel!
IT	svilupp! AND sosten!	inquin!	riscald! AND climat!	ambient! AND natural!
ES	desarroll! AND sosten!	contamin!	cambi! AND climat!	medi! AND ambient! AND natural!
DE	nachhalt! AND entwickl!	verschmutz!	klimawandel!	natur! AND umwelt!}

changement climatique is the most used in *Le Figaro*. In some cases where, after following these two steps, we were still unsure, we consulted with native speakers.

After stemming, we use Datalab procedures to develop a search within their own environment. They facilitate the use of the symbol ! (truncation), that, placed at the end of the stemmed word, includes both its root and inflections. For example, the root of the word pollution is *pollut-*, and followed by the ! sign (that would be, *pollut!*), would return all the inflections of the word such as polluted, pollution, pollutions, etc. Moreover, we use Datalab's operationalization of boolean operators, and we use the word *AND* in their search engine to select documents that contain two words. For the key terms climate change and sustainable development, we need to find articles that include both words. The term environment is a highly polysemous word that appears in very diverse semantic contexts. Therefore, we add the word nature to ensure that we are filtering our corpus efficiently for the kind of research question that we are seeking to explore.

4. Methods

Our pipeline includes five different steps. Firstly, we use the Pachinko Allocation (PA) model (Mimno et al. [1]) to discover inner topics[1] of public opinion. Secondly, we use word embeddings to calculate mean pairwise cosine similarity between words of each topic (Aletras and Stevenson [2]) - this score indicates how similar topics are. Thirdly, we use Hierarchical Cluster Analysis (Ward's linkage function) (Großwendt et al. [3]) to aggregate topics produced by all the newspapers into semantically similar global topics (separately for each epoch). Fourthly, we analyze the temporal and geographical evolution of global topics (Beykikhoshk et al. [4]). Lastly, we analyze the sentiment associated with each global topic (Hutto and Gilbert [27]).

4.1. Inner Topic Modelling

In our work, we split articles into four different epochs: (1) 1999-2003, (2) 2004-2008, (3) 2009-2013, and (4) 2014-2018. Within each epoch, we split the articles based on the newspaper they belong to. As a result, we split data into 24 subsets based on epoch and newspaper. To perform topic modeling, we use PA (Mimno et al. [1]) model on each subset of data separately. For each

[1] inner topics refer to topics that exist within a specific set of articles

subset, the PA model generates a set of topics that exist in that particular subset of articles, and we call these inner topics. These topics represent the dominating context of discourse in a particular set of articles.

We use the Pachinko Allocation model (Mimno et al. [1]) as it has several advantages over other methods, such as LDA-based models (Carlsen and Ralund [28]). It simultaneously models the correlation between words and relations between the topics. Thus, it captures topics that might exist only in a handful of articles (i.e., it is not overwhelmed by the imbalance of articles). To do so, the PA model operates on two levels: sub-topic modeling and super-topic modeling. Each generated topic is a sub-topic that the PA model has discovered. The super-topics are clusters of highly correlated sub-topics (they might share a similar vocabulary but still have distinct distribution over the vocabulary). We are particularly interested in sub-topics. However, we need to specify the number of super- and sub-topics (i.e., k^1 and k^2 parameters, see Mimno et al. [1] for more details) before we fit the PA model. Since we do not know the number of super- and sub-topics for a given dataset, we search for the most optimal set of parameters. We perform a search using Bayesian Optimisation (Head et al. [29]) that goes through many combinations of parameters and converges to the optimal solution.

The optimization algorithm requires a metric to evaluate the quality of generated topics. We use the coherence score as a proxy for the quality (Mimno et al. [30]) - it considers the words co-occurrence in a set of documents. If the co-occurrence of top-N words[2] associated with the topic is low, the coherence score would also be low (i.e., a topic does not capture any semantic relationships between the words, see Aletras and Stevenson [2]).

Algorithm 1 Inner Topic Modelling

```
for $t = 1 : \mathrm{T}$ do                                          ▷ where T is a number of epochs
    for $n = 1 : \mathrm{N}$ do                                     ▷ where N is a number of newspapers
        $\tilde{\mathbf{X}}_{t,n} = \mathrm{PreProcess}\,(\mathbf{X}_{t,n})$
        $k^1_{t,n}, k^2_{t,n} = \mathrm{BayesOptimCV}[\,\mathrm{PA}(\tilde{\mathbf{X}}_{t,n})\,]$
        $\boldsymbol{\phi}_{t,n} = \mathrm{PA}\,(\tilde{\mathbf{X}}_{t,n},\, k^1_{t,n},\, k^2_{t,n})$      ▷ where $\boldsymbol{\phi}_{t,n}$ is a set of discovered topics
    end for
end for
```

Each discovered topic in the set, $\boldsymbol{\phi}_{t,n}$, is represented as a vector, $\gamma_{t,n,d}$ that contains probabilities for each word in that topic (where $d \in D_{t,n}$ is an indicator of a topic in a set $\boldsymbol{\phi}_{t,n}$, and $D_{t,n} = k^2_{t,n}$). It provides an overview of the top words associated with the topic (i.e., to analyze the context) and a probability (i.e., to calculate the cosine similarity between different topics).

4.2. Topic Similarity

To cluster similar topics originating from different newspapers, we calculate the similarity score between each pair of topics. We do so separately per each epoch.

[2] the words that have a high probability of being in a particular topic

First, we start by estimating the similarity between inner topics. To make the multilingual topic vectors $\boldsymbol{\gamma}_{t,n,d}$ comparable, we translate words associated with each topic vector (from Italian, Spanish, German, and French) into English using the Google Translate API.

In some cases, a word translates into a phrase. To add these words to a vocabulary, we split the phrase into separate words and equally redistribute the probability associated with the phrase to these separate words. For example, according to the Google Translate *socialdemócrata* translates into *social democrat*, we pass the phrase through our pre-processing pipeline and get *social* and *democrat*. Thus, now the topic that contained *socialdemócrata* has two separate words instead of the initial term. Since we do not know anything about their probabilities and we cannot assume their independence, we equally redistribute the probabilities as $P(socialdemcrata) = P(social) + P(democrat)$, where $P(social) = P(democrat)$.

Second, as the topic similarity measure is based on word embeddings (Aletras and Stevenson [2]), we need to know all the words across every newspaper and epoch – we create a global vocabulary set, V. Global vocabulary consists of all the unique English words across every epoch and every newspaper. Now we can represent each $\boldsymbol{\gamma}_{t,n,d}$ using the global vocabulary. It might happen that a word that is mentioned in one subset of data (e.g., *cat*) might be missing from the other topic (as it was never mentioned in that particular subset of data). In that case, we update the vocabulary of the $\boldsymbol{\gamma}_{t,n,d}$ and assign a probability of 0 to all newly added words. The aligned vectors can now be used to calculate the similarities between topics.

Further, we stack all $\boldsymbol{\gamma}_{t,n,d}$ into a matrix $\boldsymbol{\Gamma} \in \mathbb{R}^{p \times v}$, where p is the total number of topic vectors (across every epoch and newspaper), and v is the length of the global vocabulary. $\boldsymbol{\Gamma}$ contains the probabilities of each word in every discovered inner topic.

Using matrix $\boldsymbol{\Gamma}$, we can extract word embeddings, i.e., numerical representations of words (Aletras and Stevenson [2]). Each word, V_i is represented as a vector where dimensions correspond to topics and values correspond to the probability of word V_i in each inner topic (i.e., the i-th column of $\boldsymbol{\Gamma}$).

We calculate the topic similarity based on the *average pairwise cosine similarity* of the N-top[3] words in each topic (Aletras and Stevenson [2]).

4.3. Global Topic Aggregation

To analyze whether newspapers share discussion points over a specific epoch, we look at the topic similarity. If multiple inner topics are similar enough, we assume they share similar contexts. The similarity is calculated between each pair of topics (produced by all newspapers over the same epoch). We cluster inner topics with the use of the Hierarchical Cluster Analysis (HCA) with Ward's linkage function (Großwendt et al. [3]).

If the similarity between topics is above a certain threshold, the HCA model clusters topics together. However, we do not have any prior knowledge of the optimal threshold value. Thus, we vary the threshold and look at the quality of the formed clusters. We use the Silhouette method (Rousseeuw [31]) as a proxy for the clustering quality. This method estimates whether topics are closer to the members of their own clusters or vice versa. For each epoch, we find a separate optimal threshold value.

[3] $N = 20$ for Sustainable Development and Climate Change; $N = 15$ for Pollution and Environment

To make clusters comparable, we create cluster representations by averaging the representations of inner topics that end up in the same cluster (i.e., the average of the corresponding rows in $\mathbf{\Gamma}$). We further refer to those as global topics.

The vectors associated with global topics are stacked into another matrix $\tilde{\mathbf{\Gamma}} \in \mathbb{R}^{g \times v}$, where g is the total number of global topics. The representation of a word, V_i, is now based on the matrix of the global topics, $\tilde{\mathbf{\Gamma}}$ (i.e., i-th columns, as in the case with the $\mathbf{\Gamma}$).

Algorithm 2 Global Topic Aggregation

```
$\mathbf{\Gamma} \in \mathbb{R}^{p \times v}$ is a matrix containing all discovered $\boldsymbol{\gamma}_{t,n,d}$ and $p$ is a total number of topics
$\mathcal{A} = \text{distance}(\mathbf{\Gamma}_i, \mathbf{\Gamma}_j) \; \forall \; i, \; j \in \{1, 2, 3, ..., p\}$    ▷ $\mathcal{A} \in \mathbb{R}^{p \times p}$ is a distance matrix
for $t = 1 : T$ do
    $\mathcal{A}^t = \text{subset}(\mathcal{A}, t)$    ▷ $\boldsymbol{A}^t$ contains only between topics of epoch $t$
    $thr_t = \text{Silhouette}(\text{Ward}(\mathcal{A}^t))$
    $\mathcal{V}_t = \text{Ward}(\mathcal{A}^t, thr_t)$    ▷ $\mathcal{V}_t$ contains sets of similar inner topics
    $\boldsymbol{\nu}_t \leftarrow \text{AverageSimilarTopics}(\mathcal{V}_\text{t}, \mathcal{V})$    ▷ $\boldsymbol{\nu}_t$ contain representations of global topics
end for
$\tilde{\mathcal{V}} = \text{stack}(\boldsymbol{\nu}_t \; \forall \; t \in \; \{1, 2..t\})$
```

4.4. Temporal evolution of topics

We are interested to see how topics change through time – do they disappear, split into multiple discourses, or stay unchanged, etc. To examine the evolution of global topics, we look at the similarities between global topics between the adjacent epochs, t and $t + 1$ (i.e., between topics of different epochs). We calculate similarities based on the above-mentioned average pairwise cosine similarity – this time, we use word embeddings from $\tilde{\mathbf{\Gamma}}$. We draw connections between the topics of the adjacent epochs if their similarity is above a certain threshold, t. Beykikhoshk et al. [4] suggests setting the threshold based on the n-th quantile of the cumulative distribution of similarity scores. When we estimate similarities between topics of every pair of t (current) and $t + 1$ (next) epochs. We order the scores, find the 90-th quantile of the cumulative distribution, and draw arrows between the pair of topics only if their similarity is higher than the 90-th quantile [4].

Due to the multi-source nature of the data, our algorithm might capture some noise. To substantiate the results, we manually inspect and correct the noisy output of the algorithm. First, we manually inspect the connection between topics if the similarity falls within the 85-th and 90-th quantile. We draw the arrow if the topics share at least one top word. Second, we remove global topics that consist only of one newspaper. We also remove topics if the values of the top ten words (in $\tilde{\mathbf{\Gamma}}$) are below 0.03[5]. We then manually inspect topics if the values of the top twenty words (in $\tilde{\mathbf{\Gamma}}$) are below 0.05. This procedure helps to remove topics lacking consistency (such as this topic with the following set of the top ten words: *leave, world, think, time, know, look, life, people, man, want*).

[4] we decided on the quantile by manually inspecting graphs

[5] we decided on the threshold by manually inspecting the results

Based on the incoming and outgoing arrows, the life of the global topics can progress in several ways: birth, evolution, split, merge, or death. It signifies how public attention and ideas are refined or transformed (Beykikhoshk et al. [4]). The birth of a topic is characterized by the absence of the incoming arrows - no topic from the previous epoch has a similar context. The death of the topic would be the reverse of this case – no topics in the future share a similar context.

If a topic has only one outgoing arrow – it evolves. The topic does not undergo a drastic change. If a topic has multiple out-coming arrows - it splits. The successors reuse similar words and context, but it also involves new words, e.g., the context surrounding these words changes. If a topic has multiple incoming connections – its ancestors merge. The context of the ancestors significantly overlaps.

4.5. Sentiment Analysis

We also calculate the sentiment score for each global topic by taking the top fifty words and passing them through the Valence Aware Dictionary and Sentiment Reasoner, VADER (Hutto and Gilbert [27]). The score signifies the sentiment associated with the context of a topic. The produced score varies between -1 (strongly negative sentiment) and 1 (strongly positive sentiment), where 0 stands for neutral. We discretize the score into seven categories (the split is based on the five equally distanced quantiles), i.e., each topic belongs to any of the seven categories.

5. Results and Discussion

After filtering our dataset using the four target key terms and the five-step pipeline, we are capable of observing both the geographic and temporal evolution of sustainability-related discourses. We are as well in a good position to observe how newspapers behave as containers of information, and, indeed, we identify consistent patterns across our four selected key terms.

We use three different metrics in the interpretation of our data analysis: diversification, attention, and geography. We are interested in observing whether topics become more diverse or simplified over time, showing incremental rates of polarity or agreement. To measure topic diversification or simplification, we use as a proxy both the labels of the topics (that we manually annotate) as well as broader semantic categories that we implement to create a second classification of topics qualitatively (and we use same colors inside the boxes of the topics to indicate equivalent subject matters). We observe that all the labels of all the newspapers across topics can be consistently grouped into six to eight semantic categories. We also quantify which topics receive the highest and lowest rates of attention by counting the number of represented newspapers as we seek to observe conflict-affinity trends over time. Finally, we are interested in measuring the geographic distribution of newspapers present in each topic, as we wonder whether it is possible to detect consistent trends of cultural diversity influencing information behaviour and, therefore, to engage with state-of-the-art that argues the existence of an Anglo-Atlantic, Southern and Central European different media tradition models (Hallin and Mancini [23]). We also scrutinize sentiment analysis behaviour to detect temporal and geographic trends.

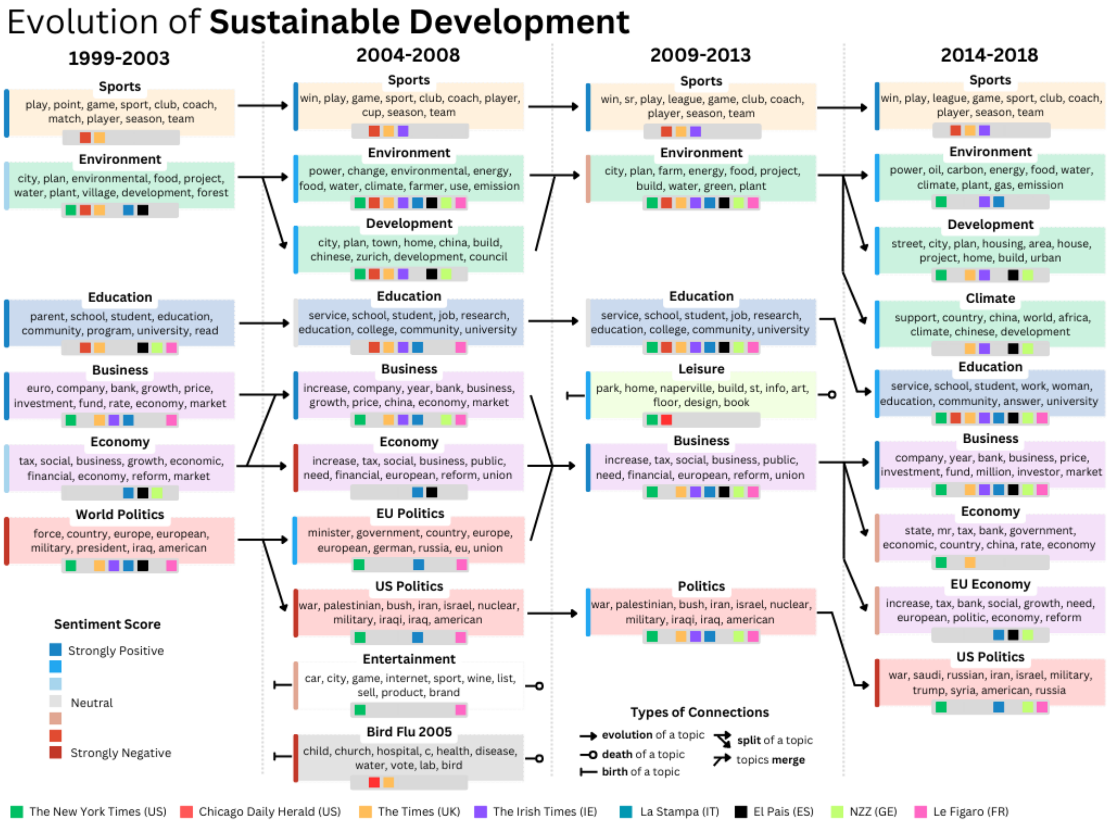

Evolution of Sustainable Development
1999-2003
2004-2008
2009-2013
2014-2018
Sports
play, point, game, sport, club, coach, match, player, season, team
Environment
city, plan, environmental, food, project, water, plant, village, development, forest
Education
parent, school, student, education, community, program, university, read
Business
euro, company, bank, growth, price, investment, fund, rate, economy, market
Economy
tax, social, business, growth, economic, financial, economy, reform, market
World Politics
force, country, europe, european, military, president, iraq, american
Sports
win, play, game, sport, club, coach, player, cup, season, team
Environment
power, change, environmental, energy, food, water, climate, farmer, use, emission
Development
city, plan, town, home, china, build, chinese, zurich, development, council
Education
service, school, student, job, research, education, college, community, university
Business
increase, company, year, bank, business, growth, price, china, economy, market
Economy
increase, tax, social, business, public, need, financial, european, reform, union
EU Politics
minister, government, country, europe, european, german, russia, eu, union
US Politics
war, palestinian, bush, iran, israel, nuclear, military, iraqi, iraq, american
Entertainment
car, city, game, internet, sport, wine, list, sell, product, brand
Bird Flu 2005
child, church, hospital, c, health, disease, water, vote, lab, bird
Sports
win, sr, play, league, game, club, coach, player, season, team
Environment
city, plan, farm, energy, food, project, build, water, green, plant
Education
service, school, student, job, research, education, college, community, university
Leisure
park, home, naperville, build, st, info, art, floor, design, book
Business
increase, tax, social, business, public, need, financial, european, reform, union
Politics
war, palestinian, bush, iran, israel, nuclear, military, iraqi, iraq, american
Sports
win, play, league, game, sport, club, coach, player, season, team
Environment
power, oil, carbon, energy, food, water, climate, plant, gas, emission
Development
street, city, plan, housing, area, house, project, home, build, urban
Climate
support, country, china, world, africa, climate, chinese, development
Education
service, school, student, work, woman, education, community, answer, university
Business
company, year, bank, business, price, investment, fund, million, investor, market
Economy
state, mr, tax, bank, government, economic, country, china, rate, economy
EU Economy
increase, tax, bank, social, growth, need, european, politic, economy, reform
US Politics
war, saudi, russian, iran, israel, military, trump, syria, american, russia
Sentiment Score
Strongly Positive
Neutral
Strongly Negative
Types of Connections
evolution of a topic
death of a topic
birth of a topic
split of a topic
topics merge
The New York Times (US)
Chicago Daily Herald (US)
The Times (UK)
The Irish Times (IE)
La Stampa (IT)
El Pais (ES)
NZZ (GE)
Le Figaro (FR)

Figure 1: Sustainable Development Discourse Evolution

Finally, we reflect on newspapers as containers of information by inspecting overall discourse behaviour across newspapers, dedicating one subsection to that purpose.

5.1. Sustainable Development

Figure 5.1 shows the historical evolution of discourses about Sustainable Development. We use boxes to represent global topics. Each box includes semantically similar words that we have grouped in the second step of our pipeline using word embeddings on the English translation of the multilingual topics that PAM outputted. We provide information of each newspaper represented in each topic using small coloured squares at the bottom of each box, matching a colour-legend that we include at the bottom of the figure. We also gather topics qualitatively into semantically similar categories (i.e., environment, politics), and we indicate this by colouring the background of each box using the same tone. We add as well sentiment analysis scores by placing a coloured bar at the left of each box, and we include a colour legend in each figure. We depict the connections that the directed graph calculates using arrows, and we explain arrow representation (types of connection) in a legend. We provide similar figures containing the same information for all our selected four key terms.

Just to provide a reading guide to understand how to interpret our figures using one column as an example, in the epoch 1999-2003, it is possible to observe six different topics (boxes): sports, environment, education, business, economy, and world politics. However, we (subjectively) group those six topics into five different semantic categories that we showcase by colouring the boxes: sports (yellow), environment (green), education (blue), business (lilac), and politics (orange). In this epoch, the topic shared by the biggest number of newspapers is world politics (*The New York Times International, The Times, The Irish Times, La Stampa, El Pais* and *Le Figaro*). The two topics that show the lowest affinity across newspapers are sports (only mentioned in *The Times* and *Chicago Daily Herald*) and economy (only mentioned in *La Stampa, El País* and *NZZ*).

And now, let's analyze the key term sustainable development using our three different criteria: diversification, attention, and geography. In terms of diversification, it is possible to observe little variation of topic labels over time, showing stability in discourse continuity over the last twenty years. That being said, it is interesting to note a progressive specialization of topics such as environment (there is a proliferation of topics in these domains from the first epoch (1999-2003) to the second one (2004-2008), followed by a simplification during the third one (2009-2013), yet increasingly fragmented during the fourth one (2014-2018)). Similarly, the business semantic category fluctuates between two topics (1999-2003, 2004-2008), to one (2009-2013), to eventually get fragmented into three as we approach contemporary times (2014-2018). Some topics are epoch specific, and do not evolve over time (2004-2008 Entertainment and Bird Flu, and 2009-2013 Leisure). The topic of politics remains relatively stable over time.

Overall, it is possible to observe a very clear stability in the evolution of sustainable environment topics over the last twenty years. While there is a clear trend of an initial diversification of topics starting in 2004-2008, it is due to the appearance of two epoch-specific topics (entertainment and Bird Flu 2005). As we have already explained, while there is a proliferation in unique topics over time (in 1999-2003 there are six of them, while in 2014-2018 there are nice), the semantic categories remain stable. So, both in 1999-2003 and in 2014-2018, discourses about sustainable development are framed under the same five semantic fields: sports, environment, education, business, and politics.

Topics that receive highest rates of attention (shared by all selected newspapers) are environment (2004-2008, 2009-2013), and education (2009-201, 2014-2018). The ones that receives the least attention are 1999-2003 sports (*The Times, Chicago Daily Herald*), 2004-2008 economy (*La Stampa, El Pais*), 2004-2008 entertainment (*The New York Times International, Le Figaro*), and Bird Flu 2005 (*Chicago Daily Herald, The Times*, 2004-2008); 2009-2013 leisure (*The New York Times International, Chicago Daily Herald*), and 2014-2018 economy (*The New York Times International, The Times*). We believe that attention could be as well be considered as a metric of polarity. The epoch that shows the highest number of low-attention topics is 2004-2008, which is precisely when discourses about sustainability fragment the most. However, we do not observe significant variations if we compare the epochs 1999-2003 and 2014-2018, as they both show similar attention rates, therefore reinforcing our observation regarding the stability of this discourse over time.

It is not possible to observe clear geographic clusters of topic affinity, with the exception of sports, that is dominated by English speaking newspapers in 1999-2003 (*The Times, Chicago*

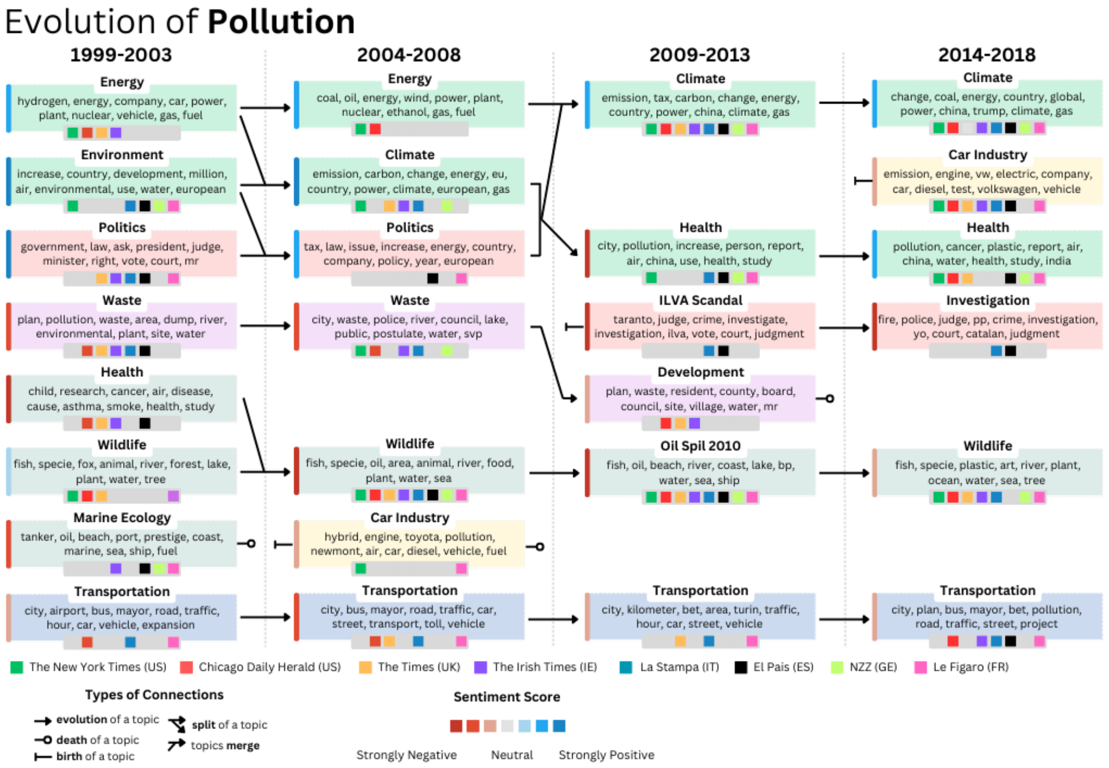

Evolution of Pollution
1999-2003
2004-2008
2009-2013
2014-2018
Energy
hydrogen, energy, company, car, power, plant, nuclear, vehicle, gas, fuel
Environment
increase, country, development, million, air, environmental, use, water, european
Politics
government, law, ask, president, judge, minister, right, vote, court, mr
Waste
plan, pollution, waste, area, dump, river, environmental, plant, site, water
Health
child, research, cancer, air, disease, cause, asthma, smoke, health, study
Wildlife
fish, specie, fox, animal, river, forest, lake, plant, water, tree
Marine Ecology
tanker, oil, beach, port, prestige, coast, marine, sea, ship, fuel
Transportation
city, airport, bus, mayor, road, traffic, hour, car, vehicle, expansion
Energy
coal, oil, energy, wind, power, plant, nuclear, ethanol, gas, fuel
Climate
emission, carbon, change, energy, eu, country, power, climate, european, gas
Politics
tax, law, issue, increase, energy, country, company, policy, year, european
Waste
city, waste, police, river, council, lake, public, postulate, water, svp
Wildlife
fish, specie, oil, area, animal, river, food, plant, water, sea
Car Industry
hybrid, engine, toyota, pollution, newmont, air, car, diesel, vehicle, fuel
Transportation
city, bus, mayor, road, traffic, car, street, transport, toll, vehicle
Climate
emission, tax, carbon, change, energy, country, power, china, climate, gas
Health
city, pollution, increase, person, report, air, china, use, health, study
ILVA Scandal
taranto, judge, crime, investigate, investigation, ilva, vote, court, judgment
Development
plan, waste, resident, county, board, council, site, village, water, mr
Oil Spil 2010
fish, oil, beach, river, coast, lake, bp, water, sea, ship
Transportation
city, kilometer, bet, area, turin, traffic, hour, car, street, vehicle
Climate
change, coal, energy, country, global, power, china, trump, climate, gas
Car Industry
emission, engine, vw, electric, company, car, diesel, test, volkswagen, vehicle
Health
pollution, cancer, plastic, report, air, china, water, health, study, india
Investigation
fire, police, judge, pp, crime, investigation, yo, court, catalan, judgment
Wildlife
fish, specie, plastic, art, river, plant, ocean, water, sea, tree
Transportation
city, plan, bus, mayor, bet, pollution, road, traffic, street, project
The New York Times (US)
Chicago Daily Herald (US)
The Times (UK)
The Irish Times (IE)
La Stampa (IT)
El Pais (ES)
NZZ (GE)
Le Figaro (FR)
Types of Connections
evolution of a topic
death of a topic
birth of a topic
split of a topic
topics merge
Sentiment Score
Strongly Negative
Neutral
Strongly Positive

Figure 2: Pollution Discourse Evolution

Daily Herald), 2004-2008/2009-2013/2014-2018 (*The Times, Chicago Daily Herald,* and *Irish Times*), Bird Flu 2005 in 2004-2008 (*The Times, Chicago Daily Herald,* leisure in 2009-2009 (*Chicago Daily Herald, The New York Times International,* and 2014-2014 economy (*The New York Times International, The Times.* There is one Southern European cluster concerned about the Economy in 2004-2008 (*La Stampa* and *El País*), showing low Sentiment Scores.

In terms of sentiments, there is a clear trend of increasing negative views in the semantic category of the economy topics as we approach contemporary times. While in the first epoch (1999-2003), scores are quite high (blue colour), in the last one (2014-2018), their evolution is negative (lower scores in the red range). The rest of the semantic categories do not show significant variations over time, highlighting once again the historic and geographic stability of sustainable development related rhetorics.

5.2. Pollution

In terms of diversification of topics, pollution shows a very clear trend of historic simplification, and, similarly to sustainable development, very little variation over time in overall discourses. In the 1999-2003 cluster, we have manually annotated eight different topics: energy, environment, politics, waste, health, wildlife, marine ecology, and transportation. In the 2014-2018, only six: climate, car industry, health, investigation, wildlife, and transportation. Furthermore, we observe how thematic categories (that, again, we showcase by colouring the topic boxes) remain very stable in the whole observational time. For example, the 1999-2003 thematic category of climate (green) remains relatively stable over time with topics such as energy and environment evolving in 2004-2008 into energy and climate, and ultimately merging into climate (2009-2013, 2014-2018). Similarly, the semantic field of politics (orange), while fragmented in two different family of topics (1999-2003 to 2004-2008 in the first place, talking about policy related issues, and 2009-2013 to 2014-2018, talking about political controversies (i.e., Ilva Scandal)), however, endures continuity in terms of representation. Waste (lilac) shows no major fluctuations, dying in 2009-2013. The semantic field of wildlife (green), appears initially highly fragmented into three different topics (health, wildlife, and marine ecology), and gets eventually simplified into just one (wildlife) in 2014-2018. Transportation (blue) appears consistently equally represented over time. Finally, the thematic category of car industry appears isolated in two epochs (2004-2008, 2014-2018).

Topics that receive the highest rates of attention are wildlife (2004-2008), climate (2009-2013), and oil spill 2010 (2009-2013). Topics that receive the lowest rates of attention are 2004-2008 energy (*The New York Times International, Chicago Daily Herald*), 2004-2008 politics (*El País, Le Figaro*), 2004-2008 car industry (*The New York Times International, Le Figaro*), 2009-2013 Ilva Scandal (*La Stampa, El País*), 2014-2018 Investigation (*La Stampa, El País*). It is possible to observe a progressive transition from polarity (topics showing low numbers of newspaper representation), to agreement (more newspapers included), as we approach contemporary times, displaying higher rates of information homogeneity in Western societies.

In terms of the geographic distribution of clusters, we observe similar patterns as in sustainable development. There seems to be some English dominated clusters, such as 1999-2003 energy (*The Times, The New York Times International, Chicago Daily Herald, Irish Times*), 2004-2008 energy (*The New York Times International, Chicago Daily Herald*); and 2009-2013 development (*Chicago Daily Herald, The Times, The Irish Times*). There is as well a clear Southern European cluster related to politics, that appears in 2004-2008 politics (*El País, Le Figaro*), with high sentiment scores; and 2009-2013 Ilva Scandal (*La Stampa, El País*), and 2014-2018 Investigation (*La Stampa, El País*). local politics (*El País, La Stampa*, and in the 2009-2013 Ilva Scandal (*El País, La Stampa*), both with low sentiment scores.

Overall, and as compared with Sustainable development, discourses appear a bit more fragmented in terms of geographic representation (there are more regional clusters), and diversity of information (there are more individual topics). However, there is a tendency towards a historic simplification as we approach contemporary times that we interpret as an increasing homogenization of views about pollution. In terms of thematic categories, and as we have just explained, there is not much variation over time and it is possible to observe how all discourses appear gathered in just six different semantic fields.

Our analysis of sentiments shows a subtle feeling of increasing optimism as we approach contemporary times. The climate semantic field (green), remains positively stable over time showing similar high scores in 1999-2003 and in 2004-2008. Politics (orange) transitions from positive (1999-2003) to negative (2014-2018). The wildlife thematic category (forest green)

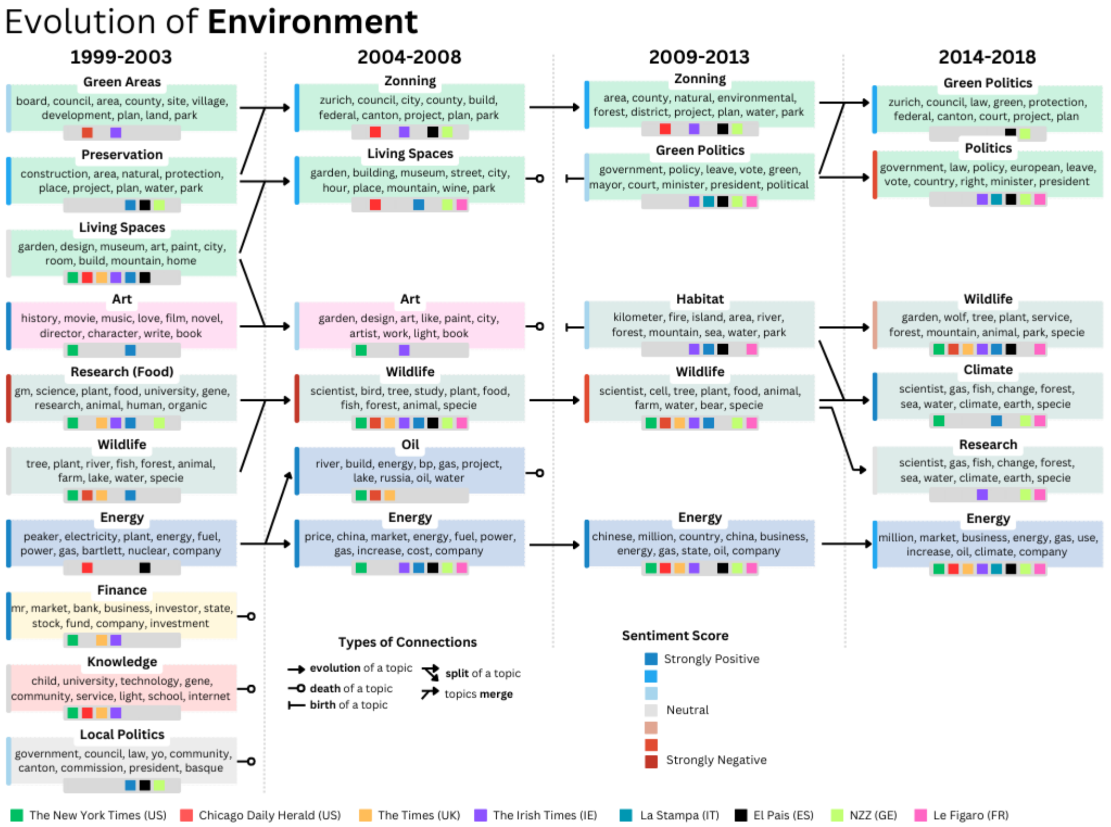

Evolution of Environment
1999-2003
2004-2008
2009-2013
2014-2018
Green Areas
board, council, area, county, site, village, development, plan, land, park
Preservation
construction, area, natural, protection, place, project, plan, water, park
Living Spaces
garden, design, museum, art, paint, city, room, build, mountain, home
Art
history, movie, music, love, film, novel, director, character, write, book
Research (Food)
gm, science, plant, food, university, gene, research, animal, human, organic
Wildlife
tree, plant, river, fish, forest, animal, farm, lake, water, specie
Energy
peaker, electricity, plant, energy, fuel, power, gas, bartlett, nuclear, company
Finance
mr, market, bank, business, investor, state, stock, fund, company, investment
Knowledge
child, university, technology, gene, community, service, light, school, internet
Local Politics
government, council, law, yo, community, canton, commission, president, basque
Zonning
zurich, council, city, county, build, federal, canton, project, plan, park
Living Spaces
garden, building, museum, street, city, hour, place, mountain, wine, park
Art
garden, design, art, like, paint, city, artist, work, light, book
Wildlife
scientist, bird, tree, study, plant, food, fish, forest, animal, specie
Oil
river, build, energy, bp, gas, project, lake, russia, oil, water
Energy
price, china, market, energy, fuel, power, gas, increase, cost, company
Zonning
area, county, natural, environmental, forest, district, project, plan, water, park
Green Politics
government, policy, leave, vote, green, mayor, court, minister, president, political
Habitat
kilometer, fire, island, area, river, forest, mountain, sea, water, park
Wildlife
scientist, cell, tree, plant, food, animal, farm, water, bear, specie
Energy
chinese, million, country, china, business, energy, gas, state, oil, company
Green Politics
zurich, council, law, green, protection, federal, canton, court, project, plan
Politics
government, law, policy, european, leave, vote, country, right, minister, president
Wildlife
garden, wolf, tree, plant, service, forest, mountain, animal, park, specie
Climate
scientist, gas, fish, change, forest, sea, water, climate, earth, specie
Research
scientist, gas, fish, change, forest, sea, water, climate, earth, specie
Energy
million, market, business, energy, gas, use, increase, oil, climate, company
Types of Connections
evolution of a topic
death of a topic
birth of a topic
split of a topic
topics merge
Sentiment Score
Strongly Positive
Neutral
Strongly Negative
The New York Times (US)
Chicago Daily Herald (US)
The Times (UK)
The Irish Times (IE)
La Stampa (IT)
El Pais (ES)
NZZ (GE)
Le Figaro (FR)

Figure 3: Environment Discourse Evolution

evolves from overall highly negative feelings in 1999-2003 to lesser ones in 2014-2018. Other isolated topics such as car industry remain negatively similar in their two respective epochs.

5.3. Environment

Environment shows very similar trends as pollution does in terms of diversification, attention, and geography. It is possible to observe a clear trend of simplification of topics over time. The first epoch (1999-2003) contains ten different topics (green areas, preservation, living spaces, art, research (food), wildlife, energy, finance, knowledge, local politics), while the last epoch (2014-2018), gets reduced to just six. In terms of semantic fields, there is a simplification ranging from six in 1999-2004 (green areas (green), art (lilac), wildlife (forest green), energy (blue), finance (yellow), knowledge (orange), local politics (grey)), to just three during 2014-2018 (green areas (green), wildlife (forest green), energy (blue)). And, overall, there are only eight different thematic categories during the whole observational time.

Topics related to environmental spaces in 1999-2003 (green areas, preservation, and living spaces), get progressively simplified into two different categories in all the rest of the epochs. The thematic field of wildlife (forest green) firstly gets simplified, and then eventually fragmented as we reach contemporary times. There is a trend of topics epoch-bounded, such as the art family (lilac), and the three 1999-2003 isolated topics of finance, knowledge, and local politics. We interpret this fragmentation as an initial high polarization across countries about environmentally related discourses.

In terms of attention, topics that receive the highest rates (all the newspapers appear represented) are wildlife (2004-2008), and energy (2014-2018). Topics that receive the lowest rates (only two newspapers represented) are 1999-2003 green areas, art, and energy; 2004-2008 art, and 2014-2018 green politics. The metric of attention therefore shows as well an initial polarization with many two-newspapers topics in the 1999-2003 epoch, progressively reaching higher rates of press agreement as we approach contemporary times.

Geographic distribution shows yet again the presence of an Anglo-speaking cluster of topics. 1999-2003 green areas (*Irish Times and Chicago Daily Herald*), finance (*The Times, The New York Times International, Irish Times*), and knowledge (*The New York Times International, Chicago Daily Herald, The Times, Irish Times*); 2004-2008 oil (*The Times, The New York Times, The Irish Times*) and art (*The New York Times, The Irish Times*). But, there is no Southern European one.

Sentiment Analysis scores show a light trend of increasing pessimism as we approach contemporary times. The green areas thematic category (green) shifts from neutral-positive sentiments in the 1999-2003 cluster to negative ones in politics related topics. The wildlife family of topics (forest green), on the other hand, becomes slightly more positive in the 2014-2018 epoch.

5.4. Climate Change

Climate change shows a simplification of topics until 2009-2013, followed by a diversification in 2014-2018. The 1999-2003 and 2004-2008 topics in the semantic field of economy (orange) disappear in 2009-2013, to re-appear in 2014-2018. Politics related topics (green) simplify in 2009-2013, to later on fragment in 2014-2018. Similarly, the thematic category of global warming (forest green), gets dramatically fragmented in 2014-2018. The leisure family of topics (lilac) follows as well this pattern. Car industry (yellow) and education (dark blue), on the other hand, remain relatively stable. Some topics are time-specific, such as the semantic category of city planning (blue). We interpret these results as a clear trend of Western societies reaching a point of international agreement in climate related topics in 2009-2013, followed by a very strong polarization in 2014-2018.

In terms of attention, topics showing the highest rates of newspapers presence include energy (2004-2008), leisure (2004-2008, 2009-2013), global warming (2009-2013), and politics (2009-2013). Topics with lowest rates of attention include 1999-2003 wine industry and car Industry, and 2014-2018 global warming, car industry, and US economy. Attention aligns with diversity, showing a higher number of low attention topics in 1999-2003 and 2014-2018 (therefore displaying polarity), and greater newspapers representation in 2009-2013 showing an agreement peak.

We note yet again an Anglo-speaking cluster of topics. 1999-2003 economy (*Irish Times, New York Times International, The Times, Chicago Daily Herald*), and sports (*Chicago Daily Herald, The New York Times International, The Times, Irish Times*), 2004-2008 education (*The Times, Chicago Daily Herald, Irish Times*), 2004-2008 Education (*The Times, Chicago Daily Herald, Irish Times*), and 2014-2018 Us Economy (*New York Times, Chicago Daily Herald*), are solely English

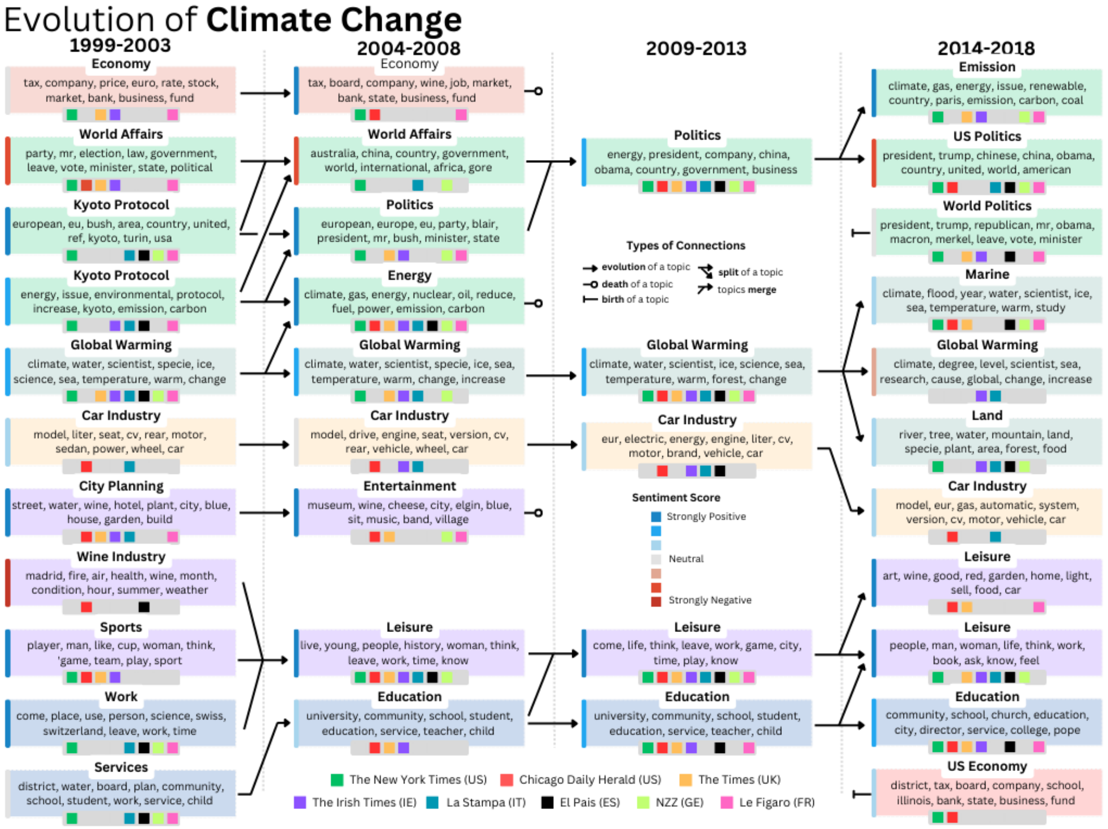

Evolution of Climate Change
1999-2003
Economy
tax, company, price, euro, rate, stock, market, bank, business, fund
World Affairs
party, mr, election, law, government, leave, vote, minister, state, political
Kyoto Protocol
european, eu, bush, area, country, united, ref, kyoto, turin, usa
Kyoto Protocol
energy, issue, environmental, protocol, increase, kyoto, emission, carbon
Global Warming
climate, water, scientist, specie, ice, science, sea, temperature, warm, change
Car Industry
model, liter, seat, cv, rear, motor, sedan, power, wheel, car
City Planning
street, water, wine, hotel, plant, city, blue, house, garden, build
Wine Industry
madrid, fire, air, health, wine, month, condition, hour, summer, weather
Sports
player, man, like, cup, woman, think, 'game, team, play, sport
Work
come, place, use, person, science, swiss, switzerland, leave, work, time
Services
district, water, board, plan, community, school, student, work, service, child
2004-2008
Economy
tax, board, company, wine, job, market, bank, state, business, fund
World Affairs
australia, china, country, government, world, international, africa, gore
Politics
european, europe, eu, party, blair, president, mr, bush, minister, state
Energy
climate, gas, energy, nuclear, oil, reduce, fuel, power, emission, carbon
Global Warming
climate, water, scientist, specie, ice, sea, temperature, warm, change, increase
Car Industry
model, drive, engine, seat, version, cv, rear, vehicle, wheel, car
Entertainment
museum, wine, cheese, city, elgin, blue, sit, music, band, village
Leisure
live, young, people, history, woman, think, leave, work, time, know
Education
university, community, school, student, education, service, teacher, child
2009-2013
Politics
energy, president, company, china, obama, country, government, business
Types of Connections
evolution of a topic
death of a topic
birth of a topic
split of a topic
topics merge
Global Warming
climate, water, scientist, ice, science, sea, temperature, warm, forest, change
Car Industry
eur, electric, energy, engine, liter, cv, motor, brand, vehicle, car
Sentiment Score
Strongly Positive
Neutral
Strongly Negative
Leisure
come, life, think, leave, work, game, city, time, play, know
Education
university, community, school, student, education, service, teacher, child
The New York Times (US)
Chicago Daily Herald (US)
The Times (UK)
The Irish Times (IE)
La Stampa (IT)
El Pais (ES)
NZZ (GE)
Le Figaro (FR)
2014-2018
Emission
climate, gas, energy, issue, renewable, country, paris, emission, carbon, coal
US Politics
president, trump, chinese, china, obama, country, united, world, american
World Politics
president, trump, republican, mr, obama, macron, merkel, leave, vote, minister
Marine
climate, flood, year, water, scientist, ice, sea, temperature, warm, study
Global Warming
climate, degree, level, scientist, sea, research, cause, global, change, increase
Land
river, tree, water, mountain, land, specie, plant, area, forest, food
Car Industry
model, eur, gas, automatic, system, version, cv, motor, vehicle, car
Leisure
art, wine, good, red, garden, home, light, sell, food, car
Leisure
people, man, woman, life, think, work, book, ask, know, feel
Education
community, school, church, education, city, director, service, college, pope
US Economy
district, tax, board, company, school, illinois, bank, state, business, fund

Figure 4: Climate Change Discourse Evolution

speaking newspapers. There is no Southern European cluster.

In terms of sentiment, we observe a very clear trend of positive emotions in 2009-2013, followed by a trend of pessimism in 2014-2018 (with special emphasis in US Politics and global warming). Therefore, in this case, sentiments line up with the three other metrics (diversification, attention, and geography), showcasing a historic break in 2009-2013.

5.5. General Trends

Overall, we observe very little variation of newspapers sustainability discourses across time of our selected four terms. While there is an array of topics that we label differently, and that, as we have explained, fluctuates in each epoch, broadly speaking it is possible to detect a very small number of thematic categories across all newspapers that could be grouped in a range of six to eight. Again, we subjectively decide how to classify topics into those broad categories. Yet we believe that our choice of semantically similar topics should not be controversial, as we think that our grouping criteria is quite self-explanatory.

For example, Sustainable development only shows eight different categories in the four epochs: 1. sports (yellow boxes), 2. environment (green boxes), 3. education (blue boxes), 4. business (lilac boxes), 5. politics (orange boxes), 6. entertainment (white box), 7. Bird Flu 2005 (grey box), 8. Leisure (lime green box).

Pollution appears framed under 6 areas: 1. energy (green boxes), 2. politics (orange boxes), 3. waste (lilac boxes), 4. wildlife (forest green boxes), 5. transportation (blue boxes), 6. car industry (yellow boxes.)

Documents containing the key terms environment and nature could be group in seven major thematic areas: 1. green areas (green boxes), 2. art (lilac boxes), 3. wildlife (forest green boxes), 4. energy (blue boxes), 5. finance (yellow box), 6. knowledge (orange box), 7. local politics (grey box).

Climate change could be classified under six major areas: 1. economy (orange boxes), 2. environmental politics (green boxes), 3. global warming (forest green boxes), 4. car industry (yellow boxes), 5. leisure (lilac boxes), 6. education (blue boxes).

Our data analysis therefore shows how newspapers content appears relatively stable over time, not showing dramatic shifts in themes represented. So, while it may be possible to observe how some topics fragment over time into an increasing number of semantically similar topics, we do not consider these as expanding information fractures but as a semantic evolution of a same thematic category. And, that is why we state that there is an overall trend of simplification of discourses as we approach contemporary times: while there may be some topic proliferation, broad semantic categories tend to get reduced, showcasing higher rates of international agreement, possibly reflecting processes of media globalization. That being said, our analysis of climate change discloses an opposite effect, highlighting a trend of agreement in 2009-2013 followed by an international public discourse crack in 2014-2018.

As we mentioned in the "Related Work" section, we seek to contribute with data-driven findings to discussions that have analyzed the agency of history, economics, and politics, in shaping different media traditions ([23]). While we do not wish to oversimplify the complexity of factors determining those different media cultures and their prevalence in today's journalistic practices, we do believe that, with the exception of climate change, our data analysis reveals an increasing homogenization of Western views as we become approach present times, although we are aware that our sample of newspapers is not big enough to make confident assertions about all Western media. That being said, these findings resonate with [23]'s views about the evolution of media systems:

> Media systems have historically been rooted in the institutions of the nation state, in part because of their close relationship to the political world. National differentiation of media systems is clearly diminishing; whether that process of convergence will stop at a certain point or continue until national differentiation becomes irrelevant we cannot yet know. (13).

At the same time, our findings reveal how sustainability related discourses have not significantly changed during the last twenty years. Even in the case of climate change, where we show a polarity breach in 2014-2018, the semantic categories are the same as in 1999-2003 (green politics, global warming, car industry, leisure, education, economy). It is important to note that our data analysis only focuses on finding semantic clusters shared by at least two newspapers.

Therefore, what we are intentionally selecting is overlapping multilingual discourses that talk about sustainability similarly: our findings are reporting points of contact in the international press. And, what we have found is that, in the majority of our selected terms, there is a solid demonstrated history of shared views among Western countries that can be traced back in time to the last twenty years. While we observe some polarity in most of the 1999-2003 epochs across terms, there is a tendency of reaching a global agreement as we become more contemporary. Furthermore, we show that the majority of the semantic categories across newspapers can be grouped in six-to-eight topics (and therefore, showing low rates of information diversity) once again reinforcing the idea that public discourses about sustainability related discourses showcase high numbers of shared views in Eurocentric societies.

Our analysis of the geographic distribution of topics empirically shows that there are no clear trends with the exception of an English speaking cluster that appears consistently over time, and a marginal southern European one related to the economy that only appears in 2/4 key terms (sustainable development and pollution). We are aware that, as our dataset lacks data from central Europe (we don´t have any newspaper representation from Germany, Austria, The Netherlands, or Belgium), as well as the Scandinavian region; it is difficult to make conclusive statements. That being said, we do use a highly diverse geographic sample that includes seven countries (Spain, France, Italy, Switzerland, UK, USA, and Ireland), and still, our clustering shows inconclusive geographic patterns.

Similarly, our study of sentiments highlights that it is not possible to notice any consistent trends across time and space. So, some terms become slightly more positive, and others a bit more negative. Consequently, we can´t detect any significant identifiable data behaviour patterns.

Finally, we do not observe significant data behaviour differences across our news outlets. While *Chicago Daily Herald* appears marginalized in discussions about politics in the key terms sustainable development, environment, and pollution (probably showing different semantics in these two areas as the rest of the newspapers), it does appear in those discussions in climate change. We do not detect any noticeable trends in *The New York Times International* signaling any differences with the rest of the newspapers.

6. Conclusion

Our article contributes to current discussions across fields about sustainability-related discourses with two major findings: a) it is possible to note a progressive homogenization of discourses across countries as we approach contemporary times showing increasing shared views across Western countries about sustainability-related topics (with the exception of climate change), and b) newspapers content in our selected key terms shows very little variation over the last twenty years, empirically demonstrating high rates of agreement in sustainability discourses historically across our selection of Western countries. We believe these are highly interesting results and that two main conclusions can be reached. Firstly, we show that even if "the West" is a multicultural space composed of a diversity of countries with different histories, economic models, religions, or languages, in terms of sustainability press discourses, it is possible to observe a rather monolithic behaviour where a common voice can be identified. And furthermore, there is a tendency to reach higher rates of agreement as we approach contemporary times, which we identify as a possible side effect of globalization processes. Secondly, we show that, in spite of recent efforts by governments worldwide, international organizations, industry, policy hubs, research funding institutions, and social activism (just to name a few), to raise awareness among civil society about the necessity of creating more sustainable societies, press discourses remain relatively stable, and therefore, not necessarily capturing recent shifts in environmental social morality. Yet, when there is strong polarity, such as in the case of climate change, newspapers do reflect discourse antagonism. Therefore, we state that newspapers discourses do not necessarily capture subtle changes in sustainability-related rhetorics, but they do reflect abrupt disagreements. Yet, as we have shown, even when there are clear trends of polarity (as it happens with climate change), newspaper content does not necessarily change: as already explained, the 1999-2003 and 2014-2018 epochs show the same semantic fields.

That being said, we wonder whether our method poses some biases in these findings. We are purposely selecting only semantic regions across countries that show the highest scores of affinity. As mentioned, in the second step of our pipeline, we use word embeddings to calculate pairwise cosine similarities of the English-translated words outputted by PAM. And in the third step, we use Ward Hierarchical Clustering to only select the highest ranked clusters. Therefore, we are intentionally eliminating topics that show low rates of agreement and that indicate polarity. Moreover, in our topic evolution graph, we are discriminating our selection of the temporal evolution of topics by only picking the ones that have the highest connection weights across epochs. So, again, we are only choosing clusters that indicate high-affinity rates across countries, leaving behind those that score low and that therefore indicate geographic disagreement on public discourses.

Yet, when there is a lot of polarity, our method is capable of uncovering it. It is possible to observe a variety of epoch-specific isolated topics such as the Ilva Scandal, the 2005 Bird Flu, the Kyoto Protocol, or the Volkswagen Car Scandal. Therefore, when newspapers content is diverse enough, our method successfully detects information variance. Moreover, our four different metrics (diversity, attention, geography, and sentiment) are capable of gradating our analysis of discourses to a rather close level. Therefore, while we may be only selecting topics that show high rates of semantic affinity over time, we have empirically demonstrated how, nevertheless, in terms of newspaper content, sustainability-related discourses have not significantly changed over the last twenty years. And, by extension, as already mentioned, we show that Western societies show relatively similar views in quite a stable way over time (even though, within that stability, there is space for disagreement, such as in climate change). Consequently, we believe that our method is capable of successfully analyzing numerically the geographic and temporal evolution of newspapers rhetorics.

Accordingly, and following this line of reasoning, we suspect that newspapers materiality may influence the composition of discourses. Newspapers are commodified information designed to be sold in a market for profit. There are certain editorial policies that most newspapers follow, such as organizing information in relatively stable sections across countries during the last twenty years (i.e., politics, economy, sports, or entertainment), number of pages, article length, font size, or publication periodicity. Therefore, we are intrigued by the thought of how the form determines the content in this specific information scenario.

Hence, future lines of work include extending our four selected words to other sustainability-related terms and expanding our current newspaper corpus to other under-represented Western regions (central Europe, the low countries, and the Scandinavian region), as well as using other computational methods to further assess the relationship between newspapers materiality and information behaviour. It would be highly interesting as well to implement a mix-methods qualitative-quantitative research approach on selected articles. Doing close readings on documents belonging to specific topics outputted by our computational analysis (i.e., comparing articles that receive low and high attention rates in terms of newspapers representation) could shed light on media environmental rhetorics and to what extent they do capture recent shifts in sustainability social morality. Finally, we are also interested to use social media (i.e., Twitter or Reddit) to test cross-platform data trends. In this article, we show how monolingual readers of their respective national press across our selected Western countries are exposed to very similar information in sustainability-related discourses. We are curious whether they may (or not) express similar views as well in social media, therefore being able to have a more well-informed judgment about the effect of the content on the form in the specific case of multilingual newspapers.

7. Acknowledgments

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie (MSC) grant agreement No 101024996. The authors would like to thank Prof. Julia Flanders for her guidance and useful advice.

References

[1] D. Mimno, W. Li, A. McCallum, Mixtures of hierarchical topics with pachinko allocation, in: Proceedings of the 24th international conference on Machine learning, 2007, pp. 633–640.

[2] N. Aletras, M. Stevenson, Measuring the similarity between automatically generated topics, in: Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, volume 2: Short Papers, 2014, pp. 22–27.

[3] A. Großwendt, H. Röglin, M. Schmidt, Analysis of ward's method, in: Proceedings of the Thirtieth Annual ACM-SIAM Symposium on Discrete Algorithms, SIAM, 2019, pp. 2939–2957.

[4] A. Beykikhoshk, O. Arandjelović, D. Phung, S. Venkatesh, Discovering topic structures of a temporally evolving document corpus, Knowledge and Information Systems 55 (2018) 599–632.

[5] United Nations, Transforming our world: The 2030 agenda for sustainable development, 2015. URL: https://wedocs.unep.org/20.500.11822/9814.

[6] J. Drucker, Sustainability and complexity: Knowledge and authority in the digital humanities, Digital Scholarship in the Humanities 36 (2021) ii86–ii94.

[7] U. K. Heise, J. Christensen, M. Niemann (Eds.), The Routledge Companion to the Environmental Humanities, Routledge, New York, NY, 2017.

[8] J. Habermas, The Structural Transformation of the Public Sphere, Cambridge, Massachusetts, 1993.

[9] E. S. Herman, N. Chomsky, Manufacturing Consent: The Political Economy of the Mass Media, Pantheon Books, New York, 2002.

[10] L. Hay, A. Duffy, R. Whitfield, Sustainability and complexity: Knowledge and authority in digital humanities, Digital Scholarship in the Humanities 32 (2021) ii86–ii94. doi:https://doi.org/10.1093/llc/fqab025.

[11] D. J. Philippon, Sustainability and the humanities: An extensive pleasure, Journal of Environmental Management 133 (2014) 232–257.

[12] S. Lenz, Is digitalization a problem solver or a fire accelerator? situating digital technologies in sustainability discourses, Social Science Information 60 (2021) 188–208. doi:doi.org/10.1177/05390184211012179.

[13] A. E. Beling, J. Vanhulst, F. Demaria, V. Rabi, A. E. Carballo, J. Pelenc, Discursive synergies for a 'great transformation' towards sustainability: Pragmatic contributions to a necessary dialogue between human development, degrowth, and buen vivir, Ecological Economics 144 (2018) 304–313.

[14] J. Godemann, G. Michelsen, Sustainability Communication: An Introduction, Springer, 2011. doi:10.1007/978-94-007-1697-1.

[15] K.-W. Brand, Sociological Perspectives on Sustainability Communication, Springer, 2011. doi:10.1007/978-94-007-1697-1.

[16] L. Kruse, Psychological Aspects of Sustainability Communication, Springer, 2011. doi:10.1007/978-94-007-1697-1.

[17] C. d. Witt, Media Theory and Sustainability Communication, Springer, 2011. doi:10.1007/978-94-007-1697-1.

[18] R. Barkemeyer, F. Figge, D. Holt, T. Hahn, What the papers say: Trends in sustainability: A comparative analysis of 115 leading national newspapers worldwide, The Journal of Corporate Citizenship 33 (2009) 69–86.

[19] R. Barkemeyer, F. Figge, D. Holt, Sustainability-related media coverage and socioeconomic development: a regional and north–south perspective, Environment and Planning C: Government and Policy 31 (2013) 716–740.

[20] A. Kumar, N. Das, A text-mining approach to the evaluation of sustainability reporting practices: Evidence from a cross-country study, Problems of Sustainable Development 16 (2021). doi:10.35784/pe.2021.1.06.

[21] D. Fischer, F. Haucke, A. Sundermann, What does the media mean by 'sustainability' or 'sustainable development'? an empirical analysis of sustainability terminology in german newspapers over two decades, Sustainable Development 25 (2017). doi:10.1002/sd.1681.

[22] V. Sebestyén, E. Domokos, J. Abonyi, Focal points for sustainable development strategies. text mining-based comparative analysis of voluntary national reviews, Journal of Environmental Management 263 (2020). doi:https://doi.org/10.1016/j.jenvman.2020.110414.

[23] D. C. Hallin, P. M. Mancini, Comparing Media Systems: Three Models of Media and Politics, Cambridge Univesity Press, Cambridge, UK, 2004.

[24] C. Sterling, Encyclopedia of Journalism, SAGE Publications, Inc., 2009. URL: https://doi.org/10.4135/9781412972048. doi:10.4135/9781412972048.

[25] R. Greenslade, The new york times introduces its new 'international edition',

The Guardian (2016). URL: https://www.theguardian.com/media/greenslade/2016/oct/11/the-new-york-times-introduces-its-new-international-edition.

[26] M. F. Porter, Snowball: A language for stemming algorithms, 2001.

[27] C. J. Hutto, E. Gilbert, Vader: A parsimonious rule-based model for sentiment analysis of social media text., in: ICWSM, The AAAI Press, 2014.

[28] H. B. Carlsen, S. Ralund, Computational grounded theory revisited: From computer-led to computer-assisted text analysis, Big Data & Society 9 (2022) 20539517221080146.

[29] T. Head, MechCoder, G. Louppe, Iaroslav Shcherbatyi, Fcharras, Zé Vinícius, Cmmalone, C. Schröder, Nel215, N. Campos, T. Young, S. Cereda, T. Fan, Rene-Rex, Kejia (KJ) Shi, J. Schwabedal, Carlosdanielcsantos, Hvass-Labs, M. Pak, SoManyUsernamesTaken, F. Callaway, L. Estève, L. Besson, M. Cherti, Karlson Pfannschmidt, F. Linzberger, C. Cauet, A. Gut, A. Mueller, A. Fabisch, Scikit-optimize/scikit-optimize: V0.5.2, 2018. URL: https://zenodo.org/record/1207017. doi:10.5281/ZENODO.1207017.

[30] D. Mimno, H. Wallach, E. Talley, M. Leenders, A. McCallum, Optimizing semantic coherence in topic models, in: Proceedings of the 2011 conference on empirical methods in natural language processing, 2011, pp. 262–272.

[31] P. J. Rousseeuw, Silhouettes: a graphical aid to the interpretation and validation of cluster analysis, Journal of Computational and Applied Mathematics 20 (1987) 53–65.

Engineering Terrorismmindedness: A Scientometric Study of the 9/11-effect on STEM Research, 1989-2022

Mats Fridlund[1] and Gustaf Nelhans[2]

[1] *Gothenburg Research Infrastructure in Digital Humanities (GRIDH), Department of Literature, History of Ideas and Religion, University of Gothenburg, Renströmsgatan 6, Gothenburg, 405 30, Sweden*

[2] *Data as Impact Lab, Swedish School of Library and Information Science, University of Borås, Borås, 501 90 Sweden*

Abstract

We study terrorism's shaping of STEM research through the development within engineering research of a 'terrorismmindedness', i.e. terrorist threat domestication through integration in research practice. This is done by a distant reading of how research in the engineering sciences is increasingly addressing terrorism-related topics. By means of an in-depth bibliometric analysis of some 3.000 terrorism-related scientific articles published 1989–2022, we construct within the subject area 'Engineering' in Web of Science its research subfield 'Terrorism Related Engineering Research'. The publications are analysed by bibliometric mapping, co-occurrence text measures and 'algorithmic historiography' using the HistCite tool. Papers cited together are mapped using VOSviewer to identify concepts and the results are clustered according to topicality, revealing the various terrorism-related research interests among engineering scientists.

Keywords

terrorism; scientometrics; bibliometrics; science and technology studies; distant reading.

1. Introduction

In April 2001 *Social Studies of Science*, the premier journal within Science and Technology Studies (STS), published a special issue on "Science in the Cold War". The historian David Hounshell concluded the issue with a commentary surveying the field. After emphasizing the importance of the Cold War he posed a question pondering the future of STS research: "If the Cold War so profoundly shaped the post-World War II world, including its intellectual outlook and research practices, what lies in store for the post-Cold War world?" [1] An answer to this question came less than half a year later with the outbreak of a new world-wide war in the form of the Global War on Terrorism (GWOT).

'9/11 changed everything', is a common phrase positing a clear before and after in the world – in all areas of society – caused by terrorism. One early scholar shortly after 9/11 pointed to the impact of the new war on technology and engineering and warned of a new "complex of military and security firms rushing to exploit the national nervous breakdown", as the new fear

> provides a powerful Keynesian multiplier. Thus the already million-strong army of low-wage security guards is expected to increase 50 per cent or more in the next decade; while video surveillance, finally beefed up to the British standard with face-recognition software, will strip the last privacy

[1] *The 7th Digital Humanities in the Nordic and Baltic Countries Conference (DHNB 2023),* March 8-10, 2023, Norway

EMAIL: mats.fridlund@gu.se (Mats Fridlund); gustaf.nelhans@hb.se (Gustaf Nelhans)

ORCID: 0000-0002-5759-0027 (Mats Fridlund); 0000-0001-5196-7148 (Gustaf Nelhans)

DHNB Publications, DHNB2023 Conference Proceedings, https://journals.uio.no/dhnbpub/issue/view/875

> from daily routine. The security regime of airport departure lounges will likely provide a template for the regulation of crowds at malls, shopping concourses, sports events, and elsewhere. Americans will be expected to express gratitude as they are scanned, frisked, imaged, tapped and interrogated ‘for their own protection‘. Venture capital will flood into avant-garde sectors developing germ-warfare sensors and threat-profile software. As the evolution of home security already illustrates, the discrete technologies of surveillance, environmental monitoring and data-processing will grow into a single integrated system. ‘Security‘, in other words, will become a full-fledged urban utility like water and power. [2]

The first major study on the impact of 9/11 on science, technology and engineering still remains to be written. Few STS-scholars studying the impact of the Cold War on research in science, technology, engineering and medicine (STEM) have applied their insights to the new war. An exception is Judith Reppy who in 2008 wondered whether the war on terrorism and the accompanied large funding and interests in bioterrorism–related R&D would lead to a new “biomedical military-industrial complex” [3] and another example is Jonathan Moreno who – inspired by earlier research by us – looked at the impact of the 9/11 attacks on research publications on neuroscience [4].

This study furthers the understanding of the impact of terrorism on science and technology and extends our earlier research where we discovered the existence of a 9/11-effect 2001–2010 on STEM research [6, 7, 8] but without any detailed analysis or in-depth studies. Here we go further through an exploration of the impact of terrorism within the area of engineering and especially within engineering research. In a wider and more general perspective, the study investigates how academic research/ers contributed to normalizing and domesticating terrorism in society through new knowledge production within engineering science that in its extension was aimed at helping citizens to better cope with terrorism in their everyday lives, i.e. how scientists assisted in engineering a wider global ‘terrorismmindedness’[5].

2. The methodologies: Computational history & quantitative STS

The core of this study is an in-depth quantitative digital history of the 9/11-effect using digital tools and resources analyzing thousands of research articles to provide a history about the impact of terrorism on STEM. In doing this it brings together two partly connected methodological developments: digital humanities within historical studies and quantitative studies of research in STS. Therefore, it should be considered both a contribution to developing digital humanities methodology with explorative bibliometric techniques, as well as to STS studies of terrorism’s impact on science and technology.

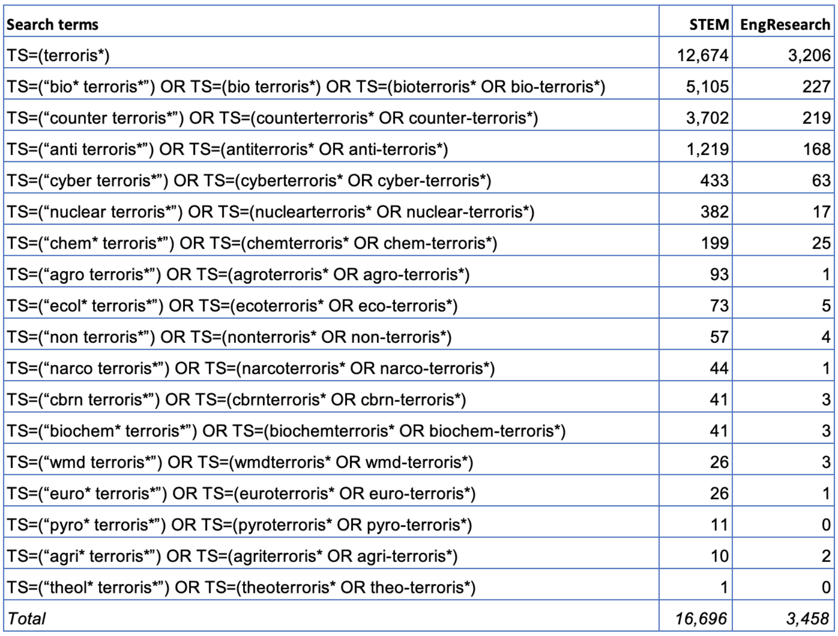
This digital history study analyzes bibliometric data taken from thousands of research articles and can through this be seen as a prototypical example of Franco Moretti’s ‘distant reading’ approach to (literary) history which he has described as where “history will quickly become very different from what it is now: it will become ‘second hand’: a patchwork of other people’s research, *without a single direct textual reading*“, [9] emphasis in original, see also [10]. In our case, distant reading of the publications means that, instead of getting information through ‘close reading’ of texts, it depends on reading and analyzing aggregated ‘metadata’ of texts: titles, author names, publication years, affiliations, keywords, and references.

The other methodological development concerns using quantitative studies of research within STS. Quantitative studies of research go back to Derek de Solla Price [11, 12, 13] pioneering work on ‘research on research’ in the 1960s. Soon bibliometric studies of scientific publications became an essential tool for such quantitative studies of research when Eugene Garfield’s Science Citation Index in the 1960s started to be used for historical and contemporary research studies. This perspective from the mid-1970s suffered a lot of critique within the nascent STS field. The key arguments for and against using quantitative data such as bibliographic information on publications and citation data came out of a ‘citation debate’ within history and sociology of science, research policy studies and STS in the mid-seventies and onward (for an in-depth analysis see Nelhans [14]). It could basically be staged as a debate between proponents advocating that citations are given to earlier research as a non-monetary reward for work done and opponents arguing that other factors also play a role when it comes to the citing of earlier literature [15]. From what could be labeled the 'institutional perspective of Sociology of Science', citations are seen as a reward in the Mertonian reward system of the norms in science [16, 17]. From this perspective one could characterize the citation as a measure of influence in some way and as indicators of scientific quality (e.g. [18, 12]. From a constructivist perspective, citations were described as indicators of rhetoric or persuasion, with its proponents denying or downplaying the utility of citations for studying research and doing history of science [19, 20, 21, 22]. In a way this debate can be seen in the light of a quality/quantification divide that went through the humanities and social sciences during the 70s and 80s and that contributed to the split of science studies at this period in time. Additionally, the question of coverage in citation databases is still an unresolved issue, where on the one hand coverage of research in peer-reviewed journal outlets is different between disciplines, but also that citation indexes predominantly cover English-language publications, which leaves out non-English language publications and especially research from the Global South.

Some of modern STS seminal scholars, such as Steve Woolgar and Bruno Latour, were involved in and influenced by the quantitative perspective at the time. The first paper in English published by Bruno Latour [23], from which traces can be seen in *Science in Action* [24], concerned the use of citations within the then nascent and later blossoming field of semiotic actor-network studies. In the 1980s, co-word analysis was developed within STS as a direct response to the scientometrics development of co-citation and bibliographic coupling methods. [25]. More recently, there has been a growing interest in utilizing bibliographical and bibliometric data within STS. For instance, van Heur et al., [26] explored the surge of the term 'ontology' in STS-related fields, while Bruno Latour and his colleagues have revisited the mapping of aggregation and emergence in research through the use of heterogeneous mappings of keywords, author names and institutional names, [27]

So it appears that both computational history and quantitative studies of research are still vibrant perspectives despite their earlier set-backs during the Cold War.

Table 1.
Terrorism in Engineering research



Search terms	**STEM**	**EngResearch**
TS=(terroris*)	12,674	3,206
TS=("bio* terroris*") OR TS=(bio terroris*) OR TS=(bioterroris* OR bio-terroris*)	5,105	227
TS=("counter terroris*") OR TS=(counterterroris* OR counter-terroris*)	3,702	219
TS=("anti terroris*") OR TS=(antiterroris* OR anti-terroris*)	1,219	168
TS=("cyber terroris*") OR TS=(cyberterroris* OR cyber-terroris*)	433	63
TS=("nuclear terroris*") OR TS=(nuclearterroris* OR nuclear-terroris*)	382	17
TS=("chem* terroris*") OR TS=(chemterroris* OR chem-terroris*)	199	25
TS=("agro terroris*") OR TS=(agroterroris* OR agro-terroris*)	93	1
TS=("ecol* terroris*") OR TS=(ecoterroris* OR eco-terroris*)	73	5
TS=("non terroris*") OR TS=(nonterroris* OR non-terroris*)	57	4
TS=("narco terroris*") OR TS=(narcoterroris* OR narco-terroris*)	44	1
TS=("cbrn terroris*") OR TS=(cbrnterroris* OR cbrn-terroris*)	41	3
TS=("biochem* terroris*") OR TS=(biochemterroris* OR biochem-terroris*)	41	3
TS=("wmd terroris*") OR TS=(wmdterroris* OR wmd-terroris*)	26	3
TS=("euro* terroris*") OR TS=(euroterroris* OR euro-terroris*)	26	1
TS=("pyro* terroris*") OR TS=(pyroterroris* OR pyro-terroris*)	11	0
TS=("agri* terroris*") OR TS=(agriterroris* OR agri-terroris*)	10	2
TS=("theol* terroris*") OR TS=(theoterroris* OR theo-terroris*)	1	0
Total	*16,696*	*3,458*

3. The phenomenon: The 9/11-effect on STEM research

In this study, the identification of engineering research pertaining to the topic of 'terrorism' is based on a set of scientific articles published between the years 1989 and 2022 identified in Clarivate's *Science Citation Index Expanded* (SCI-E) and *Conference Proceedings Citation Index – Science* (CPCI-S) which are here taken together referred to as *Web of Science* (WoS). It is to be noted that WoS does not contain all research publications and is not complete in any major way. Instead, according to what it deems the ‘most relevant’ scientific publications – mainly based on citation metrics – it indexes a broad range of journals and conference proceedings which means that appropriate and possibly significant literature (including monographs and “gray literature”) not covered by WoS will be missing from our analysis. Furthermore, there is a Western bias of journals indexed by WoS, making authors from the global south disproportionately less featured in their coverage. [28] The articles selected for analysis were published in journals classified under the ‘Engineering’ subject area in the Clarivate databases. Only articles containing the term **terrorist** within their title, abstract or author-generated keywords were included for analysis.

Through analysis and use of bibliometric methods, we delineate the emergence of a research field and a research community of what we have called *Terrorism Related Engineering Research* (TRER). However, as we are the ones defining this field, it is possible that the researchers within it may not necessarily recognize it as a distinct field or community. Nevertheless, it is an actual research field in that the research is unified in its inferred (and often implied) relevance to terrorism.

The search criteria for this study included various combinations of terms related to the term ‘terrorism’ in titles, abstracts, and author-generated keywords. As the WoS interface does not allow truncation of search terms at the beginning of a word, we have manually identified compound forms of the concept using possible prefixes and hyphen-based variants. We have extensively browsed the literature and consulted available dictionaries to identify relevant variants that were used in the searches. Table 1 provides an overview of search terms and total number of papers found in WoS, as a whole and refined by research area and publication type as ‘Engineering research’.

In all, 3,458 terrorism-related articles were found for the engineering research set (EngResearch) and as such comprise all the TRER publications. A broader set, comprising 16,696 articles within the Science-related databases were also retrieved for reference (STEM). First, we will describe this set. Figure 1:(a-b) displays the annual number of published articles within each set. Part (a) shows the total number of scientific articles for each set using the same scale. As a green curve, the total number of indexed articles in SCI-E and CPCI–S is shown. In contrast, part (b) employs different scales to underscore yearly differences and similarities.

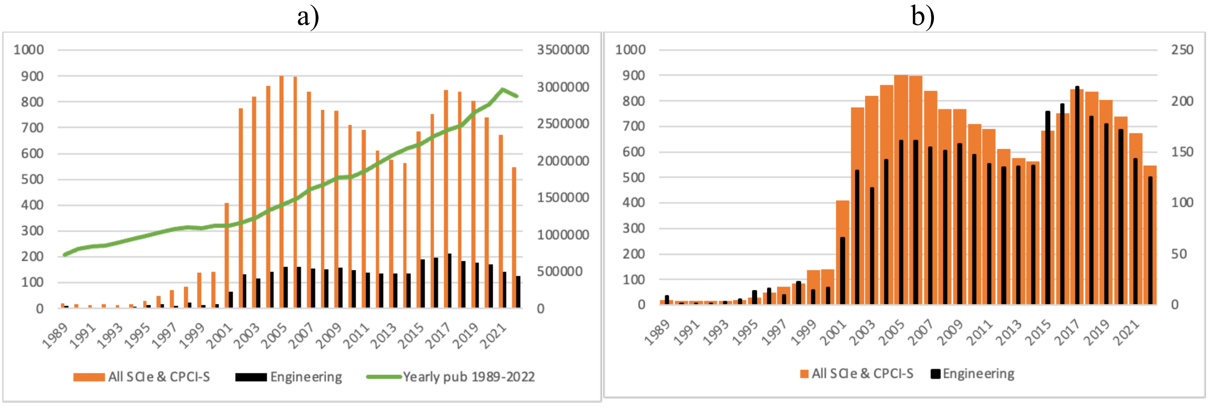
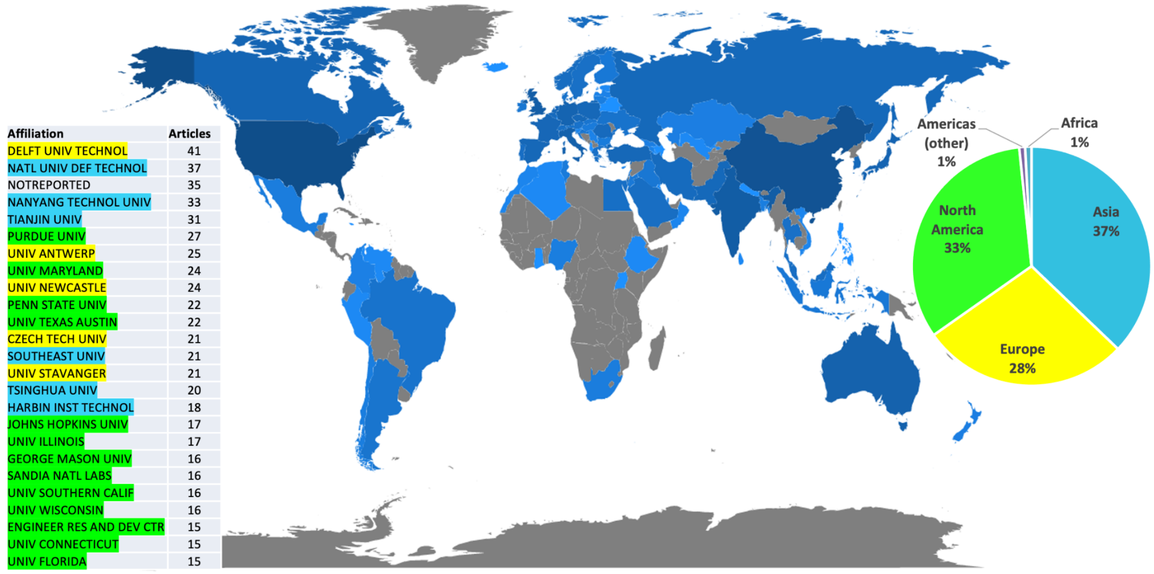


Figure 1: Yearly development for the SCI-e/CPCI-S scientific papers versus the TRER engineering papers. (a). science (orange) and engineering research (black) with the same scale and (b). both sets with different scales.

The first graph (a) of STEM research indicates a gradual growth from 1989 until 2001, with less than 200 articles before 2001. Subsequently, there was a phase of rapid growth from 2001 to 2002, and the trend continued until 2006 peaking at 900 publications. That is followed by a decrease until 2014. Then again, there is a second growth 2015-2018, followed by yet another decrease until the last year of the study. There are at least three noteworthy points to consider. First, the growth of WoS is largely linear during the study period, suggesting that variations in indexing cannot account for the observed differences in the graphs. Second, while the growth may be slightly overestimated due to the linear expansion of the full database, the decline is even more pronounced than depicted, given that WoS has been increasing its number of articles by 4 % annually, as per WoS data. Third, it is important to investigate the factors responsible for the decline in 2006–13, as well as the second peak, along with the accompanying growth and decline. The initial decline could be due to a waning interest in terrorism-related research among scholars, but it is also plausible that the decline is a result of ‘obliteration by incorporation,’ where researchers could be conducting TRER without using specific terrorism-related terms. This possibility is especially likely given the increasing usage after 2001 of ‘homeland security’ as another key terrorism-related term. And especially after November 2002 with the founding in the USA of its Department of Homeland Security.

The second graph (b) displays the occurrence of terrorism-related terms in engineering research (TRER), which exhibits a very similar trend to STEM but at a lower scale. Notably, TRER represents a substantial part – between 20% to 50% – of all terrorism related research in STEM before 1998. However, it is important to be cautious of variations, as even though they may be statistically significant they could be influenced by local events, as well as special issues or conferences. Nevertheless, the ENG and STEM trends are comparable. We have also conducted a similar search for STEM-research using ‘nuclear war’ (not shown) which reveals a decline in the Cold War research from the end of the Cold War in 1985 until 2001 and an accompanying waning of Cold War ‘nuclearmindedness’.

That the trends are very similar becomes very clear when looking at graphs (a) and (b) where in (a) STEM and ENG are shown in the same scale and where TRER stands for about 20% of the total hits of STEM. In (b) the two graphs are shown in different scales which shows a very close match in trends regarding increase and decline. A possible difference is that the decline in ENG might be somewhat later in setting in than in STEM depending on whether the 2009 peak for ENG is an anomaly or to be seen as a representative of the actual interest among researchers.



Affiliation	Articles
DELFT UNIV TECHNOL	41
NATL UNIV DEF TECHNOL	37
NOTREPORTED	35
NANYANG TECHNOL UNIV	33
TIANJIN UNIV	31
PURDUE UNIV	27
UNIV ANTWERP	25
UNIV MARYLAND	24
UNIV NEWCASTLE	24
PENN STATE UNIV	22
UNIV TEXAS AUSTIN	22
CZECH TECH UNIV	21
SOUTHEAST UNIV	21
UNIV STAVANGER	21
TSINGHUA UNIV	20
HARBIN INST TECHNOL	18
JOHNS HOPKINS UNIV	17
UNIV ILLINOIS	17
GEORGE MASON UNIV	16
SANDIA NATL LABS	16
UNIV SOUTHERN CALIF	16
UNIV WISCONSIN	16
ENGINEER RES AND DEV CTR	15
UNIV CONNECTICUT	15
UNIV FLORIDA	15

Figure 2: Geographical and organizational information about the TRER research.

Following this we have looked at the country distribution among the listed TRER-authors (Figure 2) However, this is the number of authors rather than publications which skews the data somewhat as many publications are multi-authored and multinational. What is very apparent and rather expected from the graph and the tables is the dominance of the USA with almost twice the number of authors than the second most prolific country, China. When it comes to continents, Asia (37%) actually surpasses North America (33%) in the share of authors, followed by Europe at 28%. Worth mentioning are the very few authors from Africa (except for Egypt and South Africa). The last 10 years have seen a big increase in the share of Asian authors, who stood only for 18% 1989–2013.

Regarding the table of listed institutions there are similar patterns of US dominance. There is worth mentioning some of the results illustrating the diversity of the institutions involved. One is that besides the number of US universities we also have a Norwegian (University of Stavanger) and an Argentinian institution among the listed. Furthermore, we also see private companies such as the government and military contractor Booz Allen Hamilton among the institutions, something which might indicate connections to an emerging military-industrial-academic complex centered on terrorism-related research.

4. Distant readings: Bibliometric analysis of the 9/11-effect

Up to here we have only discussed bibliographic metadata of the publications relating to the TRER set directly retrieved from the citation index. This was discussed primarily from a quantitative aspect. In this section we will start to dig a bit deeper in the data by introducing citation measures to identify aspects of the publications that are not found by ranking based on quantity of publication.

By analyzing the temporal patterns of the use of distinct terms and in analyzing different variables such as the subject space (based on the names of the publication sources), the origin of the research in terms of organization and national distribution it is possible to map the material and to identify relevant trends that then could be examined using other quantitative methods or in a deeper qualitative case studies where specific instances that are identified could be focused on. These bibliometric tools then act as a focusing lens which highlights (and to some extent suggests interpretations of) relevant areas in the research material that could be more specifically focused on.

Two different basic citation scores will be used here. One is the traditional measure that is found in WoS for a published paper when retrieved in the database. In this study, this will be described as the Global Citation Score (GCS) for the entity (paper, author or source journal/conference) discussed. This is also referred to as the external impact measure (EIM), since it calculates the amount of influence that each entity has performed in the whole of WoS. Additionally, a Local Citation Score (LCS) will be introduced that pertains to the number of citations each entity has received *within* the set of 3,458 TRER-publications. This is regarded as an internal impact measure (IIM) of its relevance and impact specifically on the constructed research field (TRER), since it measures the amount of influence exercised on the literature within the set of papers that has been created for bibliometric analysis.

In the next section, the publications will be bibliometrically mapped according to topical properties, (co-citations on journal level, and cooccurrence of noun phrases within the titles and abstracts) where journals often cited together will be found to be clustered more closely together in the visualization, thus suggesting them having more in common than other papers or authors cited by different literatures, that are not found close to the cluster. In the same way, co-occurring phrases form clusters that could be visually analyzed. The resulting visualizations were investigated both quantitatively and qualitatively, where key publications identified in a specific cluster in the visualization were selected for close readings to elucidate the qualitative historical effects of the 9/11-effect on engineering research.

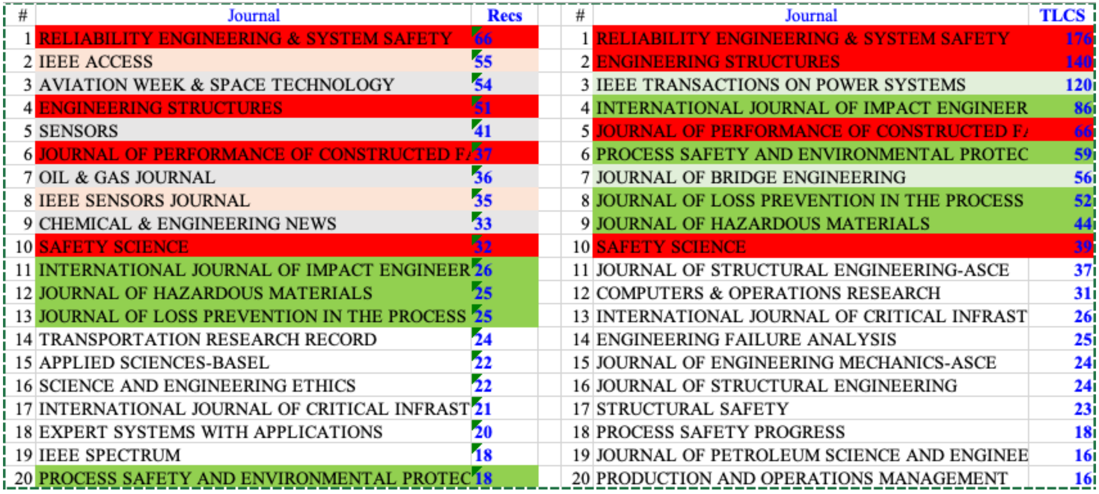
4.1 TRER publication forums

The top publication forums according to most terrorism-related publications in journal/proceeding is shown in table 2. To the right the top relevant forums are ranked according to most highly-cited/highest impact within terrorism-related research, shown as Total LCS (TLCS) score. A few of the top ten journals (i.e. #3, #7, #9) according to the number of articles published are not very research-relevant. Additionally, several of the other sources with most publications have very few, or zero citations. According to the titles and the number of received citations they have acquired they appear to be journals directed towards more applied professional communities rather than scientific research communities. Also, one journal, *Sensors*, is published by a publisher that sometimes is criticized for problematic publishing practices [29].

It is therefore more relevant to focus on TLCS – 'Internal Impact Measure', IIM. Turning to the right-hand side, it could be found that the list comprises regular scientific journals and conference proceedings. Some journals were only found on the left-hand side (indicated by gray color). Journals in red and green color depict those who publish frequently and have a high IIM. Light red are those who publish frequently but have a lower IIM, while light green journals have a relatively high IIM, but a low number of publications Although a large overlap occur, many of the highest journals outside of the top ten are not found among the most frequently publishing journals, indicating that journals don't have to publish a large number of articles to become relevant to a research field.

Topic wise the areas standing out are civil engineering and construction, chemical engineering, power systems (highest) and more generally 'process and operations management'. This is focusing on areas compared to the larger diversity among the highest publication forum to the left. What we start to see here is what specific topic areas that are of central interest related to terrorism in engineering, although we don’t yet see many details of the research.

Table 2: Top publication forum according to most TRER publications in journal/proceeding.



#	Journal	Recs	#	Journal	TLCS
1	RELIABILITY ENGINEERING & SYSTEM SAFETY	66	1	RELIABILITY ENGINEERING & SYSTEM SAFETY	176
2	IEEE ACCESS	55	2	ENGINEERING STRUCTURES	140
3	AVIATION WEEK & SPACE TECHNOLOGY	54	3	IEEE TRANSACTIONS ON POWER SYSTEMS	120
4	ENGINEERING STRUCTURES	51	4	INTERNATIONAL JOURNAL OF IMPACT ENGINEER	86
5	SENSORS	41	5	JOURNAL OF PERFORMANCE OF CONSTRUCTED F/	66
6	JOURNAL OF PERFORMANCE OF CONSTRUCTED F/	37	6	PROCESS SAFETY AND ENVIRONMENTAL PROTEC	59
7	OIL & GAS JOURNAL	36	7	JOURNAL OF BRIDGE ENGINEERING	56
8	IEEE SENSORS JOURNAL	35	8	JOURNAL OF LOSS PREVENTION IN THE PROCESS	52
9	CHEMICAL & ENGINEERING NEWS	33	9	JOURNAL OF HAZARDOUS MATERIALS	44
10	SAFETY SCIENCE	32	10	SAFETY SCIENCE	39
11	INTERNATIONAL JOURNAL OF IMPACT ENGINEER	26	11	JOURNAL OF STRUCTURAL ENGINEERING-ASCE	37
12	JOURNAL OF HAZARDOUS MATERIALS	25	12	COMPUTERS & OPERATIONS RESEARCH	31
13	JOURNAL OF LOSS PREVENTION IN THE PROCESS	25	13	INTERNATIONAL JOURNAL OF CRITICAL INFRAST	26
14	TRANSPORTATION RESEARCH RECORD	24	14	ENGINEERING FAILURE ANALYSIS	25
15	APPLIED SCIENCES-BASEL	22	15	JOURNAL OF ENGINEERING MECHANICS-ASCE	24
16	SCIENCE AND ENGINEERING ETHICS	22	16	JOURNAL OF STRUCTURAL ENGINEERING	24
17	INTERNATIONAL JOURNAL OF CRITICAL INFRAST	21	17	STRUCTURAL SAFETY	23
18	EXPERT SYSTEMS WITH APPLICATIONS	20	18	PROCESS SAFETY PROGRESS	18
19	IEEE SPECTRUM	18	19	JOURNAL OF PETROLEUM SCIENCE AND ENGINEE	16
20	PROCESS SAFETY AND ENVIRONMENTAL PROTEC	18	20	PRODUCTION AND OPERATIONS MANAGEMENT	16

Also we see three groupings of the IIM with the three top journals above 100 local citations, and then a grouping of five journals with more than fifty citations each. It is noteworthy that although the two most internally relevant journals also are the most popular journals to publish within the terrorism engineering research community. Many of the subsequent journals on the cited list are not among the most frequent in terms of published articles. What this tells us is that the most interesting and important research within the terrorism research field is to be found on the right side. This is important for further in-depth research. What complicates matters is that researchers citing practices are somewhat biased toward citing the same journal that they publish in, since it can be assumed that the readers of the article also have access to earlier issues of the same journal [30].

4.2 Leading TRER authors

Like for the journals the most productive researchers are not necessarily the most relevant or influential among the research community. In Table 3, the top 20 authors are shown sorted based on three criteria: amount of published papers in the set, external impact (TGCS) and internal impact (TLCS). Here, external impact is thought of as a means to say something about what TRER that is relevant to the outside research community.

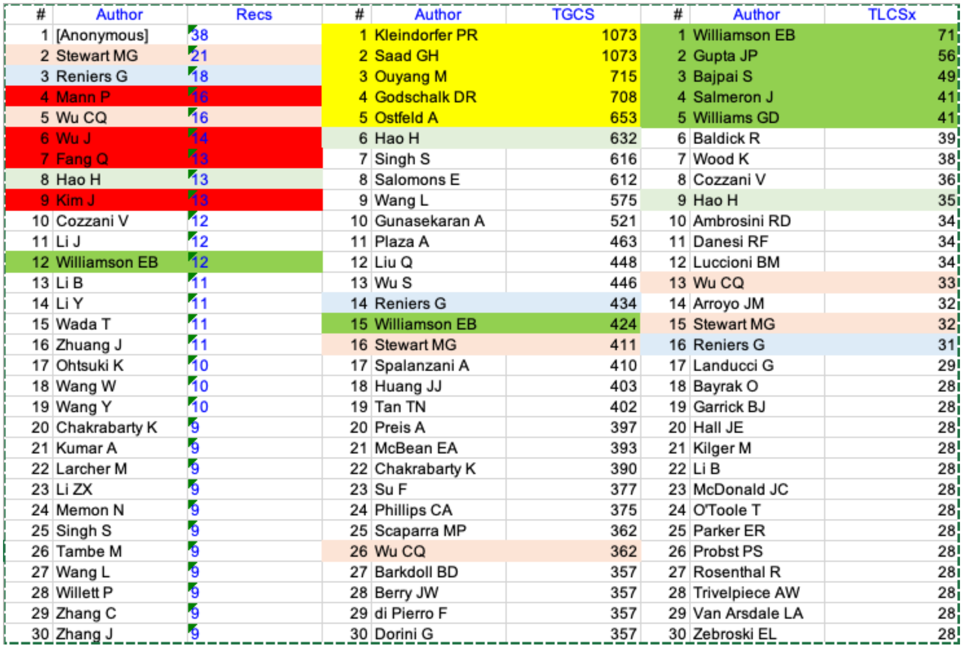
One example is the paper by Kleindorfer & Saad, the highest cited and most influential paper in the external community, whose article is about supply-chain management but whose abstract only includes ‘terrorism’ as an aside remark, as seen from its bibliographical data and abstract extract:

> Kleindorfer PR (Kleindorfer, PR); Saad GH (Saad, GH), ”Managing disruption risks in supply chains”, PRODUCTION AND OPERATIONS MANAGEMENT 14 (1): 53-68, 2005
>
> Abstract: There are two broad categories of risk affecting supply chain design and management: (1) risks arising from the problems of coordinating supply and demand, and (2) risks arising from disruptions to normal activities. This paper is concerned with the second category of risks, which may arise from natural disasters, from strikes and economic disruptions, and from acts of purposeful agents***, including terrorists***. [Our emphasis] The paper provides a conceptual framework that reflects the joint activities of risk assessment and risk mitigation that are [...]

Arguably, what this could indicate is research that tries to give itself more contemporary relevance by adding terrorism related terms to its abstract. It should be noted that this is not a judgment of the intentions of the authors of this paper, but a suggestion in need of further studies.

The most influential article in TRER is by Salmeron et al on electrical grid security under terrorist threat, which is a research area related to the most influential researcher Williamson, who publishes research on Bridges/Construction (Performance of Bridge Columns Subjected to Blast Loads. I: Experimental Program).

Table 3: Authors ranked based on No. of published papers (Recs), external impact (TGCS) and internal impact (TLCS)



#	Author	Recs	#	Author	TGCS	#	Author	TLCSx
1	[Anonymous]	38	1	Kleindorfer PR	1073	1	Williamson EB	71
2	Stewart MG	21	2	Saad GH	1073	2	Gupta JP	56
3	Reniers G	18	3	Ouyang M	715	3	Bajpai S	49
4	Mann P	16	4	Godschalk DR	708	4	Salmeron J	41
5	Wu CQ	16	5	Ostfeld A	653	5	Williams GD	41
6	Wu J	14	6	Hao H	632	6	Baldick R	39
7	Fang Q	13	7	Singh S	616	7	Wood K	38
8	Hao H	13	8	Salomons E	612	8	Cozzani V	36
9	Kim J	13	9	Wang L	575	9	Hao H	35
10	Cozzani V	12	10	Gunasekaran A	521	10	Ambrosini RD	34
11	Li J	12	11	Plaza A	463	11	Danesi RF	34
12	Williamson EB	12	12	Liu Q	448	12	Luccioni BM	34
13	Li B	11	13	Wu S	446	13	Wu CQ	33
14	Li Y	11	14	Reniers G	434	14	Arroyo JM	32
15	Wada T	11	15	Williamson EB	424	15	Stewart MG	32
16	Zhuang J	11	16	Stewart MG	411	16	Reniers G	31
17	Ohtsuki K	10	17	Spalanzani A	410	17	Landucci G	29
18	Wang W	10	18	Huang JJ	403	18	Bayrak O	28
19	Wang Y	10	19	Tan TN	402	19	Garrick BJ	28
20	Chakrabarty K	9	20	Preis A	397	20	Hall JE	28
21	Kumar A	9	21	McBean EA	393	21	Kilger M	28
22	Larcher M	9	22	Chakrabarty K	390	22	Li B	28
23	Li ZX	9	23	Su F	377	23	McDonald JC	28
24	Memon N	9	24	Phillips CA	375	24	O'Toole T	28
25	Singh S	9	25	Scaparra MP	362	25	Parker ER	28
26	Tambe M	9	26	Wu CQ	362	26	Probst PS	28
27	Wang L	9	27	Barkdoll BD	357	27	Rosenthal R	28
28	Willett P	9	28	Berry JW	357	28	Trivelpiece AW	28
29	Zhang C	9	29	di Pierro F	357	29	Van Arsdale LA	28
30	Zhang J	9	30	Dorini G	357	30	Zebroski EL	28

On the other hand, we also find several very relevant researchers and if we take a look at some of the highest internally influential researchers, such as Gupta (6 pubs), Bajpal (5 pubs) (co-authoring on the topic of risk assessment in oil and gas, as well as chemical industries., and Salmeron (2 pubs), we find that they are not very highly productive, since these authors are not found among the top 30 publishing authors in the table to the left.

4.3 Cited references

Another way of identifying relevant topics in the data is by way of the historiograph [31, 32, 33]. Algorithmic historiography, a concept invented by Garfield [34], was first put to use in tracing the history of DNA through computational methods. Citation data, Garfield argued, could help trace the lineage of history of science, to indicate on "whose shoulders' ' [35] researchers stand on, and who subsequently extend and further develop these lines of thought. This idea lay dormant for three decades, until put to use in the HistCite software in the 2000s, thanks to developments in computer power and hyperlink technology [32, 36].

This is a way of visualizing the citation network as a tree structure or as a family tree (Figure 3). What we have here is a diagram, which as its vertical axis has the years of publication. Every node represents a specific paper, where the size of the circle shows its relative frequency of citations. Each line connecting nodes represents that the subsequent node refers to the paper above. The numbers within the nodes are an id# that can be used to find the specific paper in a legend database for the map. The horizontal axis is laid (without order) to increase readability.

The internal influence map shows the 10 most cited articles and those TRER researchers influenced by that research. Here we could visually identify clusters of articles referencing each other. By browsing through the publication titles (as identified from the id#) we can assume the topic of the research in each citation cluster. The graph has a horizontal line indicating the year 2001 which indicates that only one paper published before that year was found to be of high internal relevance in the set. Among the 71 articles that have received more than 1 citations locally in the set only one other was published before

2001. This is truly remarkable as citations are cumulative and that older articles in general tend to be more highly cited on average due to the fact that citations accumulate over time.

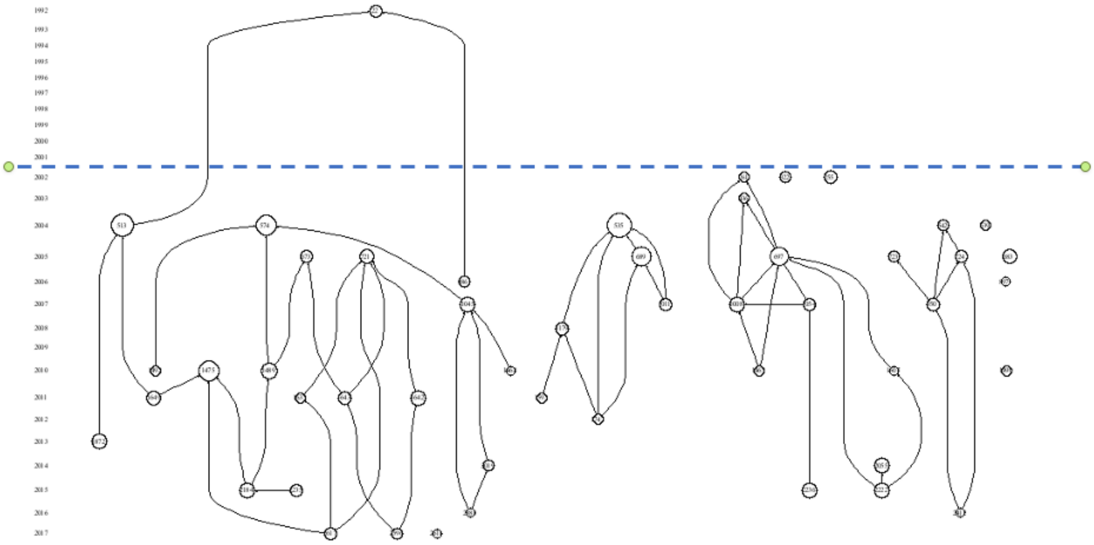
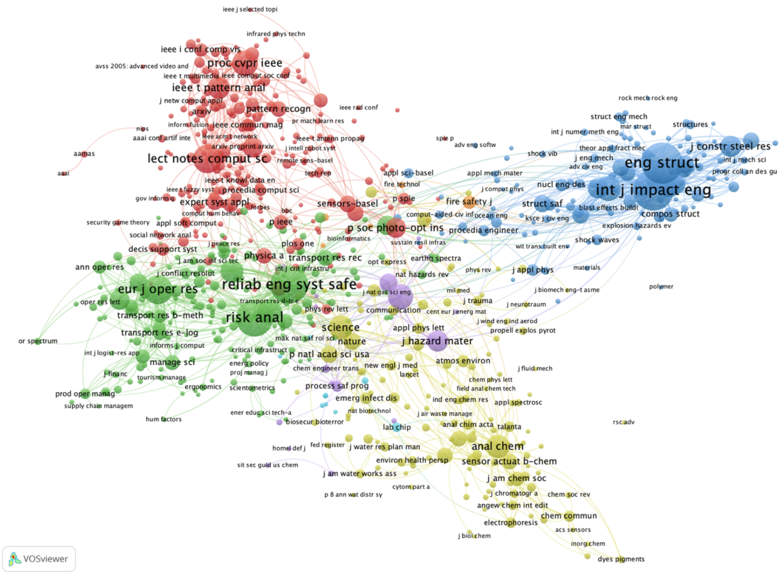


Figure 3: Internal influence map (historiograph) of the ten most cited articles within the set, together with the referring papers. The clusters from left to right concerns buildings and explosions, infrastructure planning, industrial safety, buildings and materials. The 1992 article is titled "The protection of buildings against terrorism and disorder".

The first (leftmost) clusters relate to buildings and impact blasts. One of the early highly cited articles is #255: 'Why did the World Trade Center collapse? - Simple analysis'. The first cluster starts already in 1992 with an article (#22) on "The Protection of Buildings Against Terrorism and Disorder", which was taken up by a 2004 paper on 'Analysis of Building Collapse under Blast Loads', among others (513). The next cluster, starting with #574 'Confronting the Risks of Terrorism: Making the Right Decisions' and #697: 'Site Security for Chemical Process Industries' could be referred to as pertaining to decision making, connected to protecting critical infrastructure. Lastly, there is a larger cluster consisting of articles on modeling in relation to electric grid security with the aforementioned authors of internal high impact such as Salmeron et al (#535) and Arroyo (#689).

Another way of looking at topicality is by way of looking at the research fronts of research based on measures of co-cited sources by the set of articles that the bibliometric study relates to. With this graph search method, it is possible to find potentially terrorism-related and/or relevant research that does not use the term terrorism. In Figure 4 journals cited together in the TRER set are mapped using the software VOSviewer [37] based on the degree of relatedness between the journals. The closer two journals are plotted, the more often they are cited together in the TRER papers in our set. The size of the nodes relates to the number of times they are cited. The results are clustered according to topicality. Here four main clusters, comprising four branches could be discerned that roughly match the result from the internal influence map (historiograph) in Figure 4 above, except for one important area. Electrical grid security is not found, while the red area on the top instead relates to the computational modeling and recognition using various sensors, the green to decision-making purposes and the blue to buildings and impacts blasts. The yellow cluster refers to biochemical research that while not having any highly cited papers within the set (thus not found in Table 3) still make a highly relevant topic of TRER research. The yellow cluster is a candidate for a topic that could be valuable to investigate, but from the journal titles it is not possible to discern exactly what the papers are about. What is interesting here is that the cited objects do not just refer to journals in WoS indexed journals. Instead, any publication that is found in the reference lists of the TRER papers is amenable to be cited (e.g., books, reports and non-WoS-indexed journal articles), increasing the range of the bibliometric analysis outside of the limitations of the WoS citation index coverage.



Source	Citations
eng struct	792
int j impact eng	771
reliab eng syst safe	667
risk anal	626
eur j oper res	395
lect notes comput sc	376
proc cvpr ieee	312
j perform constr fac	298
j loss prevent proc	292
science	276
safety sci	266
anal chem	258
ieee t power syst	241
ieee t pattern anal	234
j hazard mater	210
j constr steel res	208
j struct eng	202
ieee access	195
j struct eng-asce	195

Figure 4. The intellectual base: Co-cited journals with at least two citations from the TRER set. Of 23,339 cited references 752 cited sources with ≥10 citations are shown.

5. Diachrony & dendrograms: Changing TRER research topics

Another complementary way of graphing the literature is through what we call *scientometric diachrony*, through studying the terms and concepts in scientific publications' titles and abstracts and their change over time. This was done using the VOSviewer software to distant read the *contents* of TRER publications (rather than their bibliographic data). The algorithm not only connects terms found in the analyzed texts but also uses a method based on a linguistic database that connects noun-phrases and phrases containing an adjective before the noun into concepts that are shown. Furthermore, it uses the TF/IDF technique to identify the most relevant noun phrases. It does so by weighting concepts based on their occurrences in the texts. Phrases that are commonly found across many texts, such as *paper, interesting result,* or *new study*, are weighted low, while specific concepts that are only found in certain contexts, such as *bioterrorism, drone* and *toxin*, are weighted higher [38].

The co-word maps produced with this method graphs significant terms and how they are related to each other within and across the set of texts that are analyzed. This could be regarded as a means of distant reading of the texts that are included in the set. There is no practical way of manually reading the whole set of papers, but using indicator-based methods such as these, we argue that there is potentiality in getting insights in the literature at an aggregate level that could not be found without these techniques. It is also argued that these methods do not replace traditional close reading of texts, but that they amend the methodological arsenal and could be used to find different insights than those gathered by close reading.

We also used a second method for identifying relevant terms by subjecting the data to factorial analysis in Bibliometrix [39] that uses multiple correspondence analysis to elicit key terms and their relationships in a dendrogram that displays the relation between terms in a hierarchical manner. The advantage of this method is that it elicits few, but distinct related terms that describe the data. By triangulating between the co-word maps and the topic dendrograms, we were able to identify more subtle details of the material than either of the methods could do on their own.

The difference between the two visualizations is that the VOS co-word map generates many more terms than the topic dendrogram but its terms are overlayered and thus terms that might be better at indicating the overall cluster topic might be covered by the foremost terms, while the fewer dendrogram topic terms are results of terms aggregated at a higher level which takes away specificities and nuances of the terms aggregated. This makes it preferable to use the methods complementary and at times rely more on one rather than the other method.

Here, it is important to note that although these methods are highly quantitative in their nature, employing many levels of statistics and exact samples, there is a clear qualitative stance towards it. The clusters that are developed are not automatically given, but we as researchers make active choices regarding cut off points that lead to variations in the number of clusters that are mapped and the level of detail in the analysis. In doing these analyses the researchers work iteratively, moving back and forth between results and interpretations to close in on the final results to be presented and discussed.

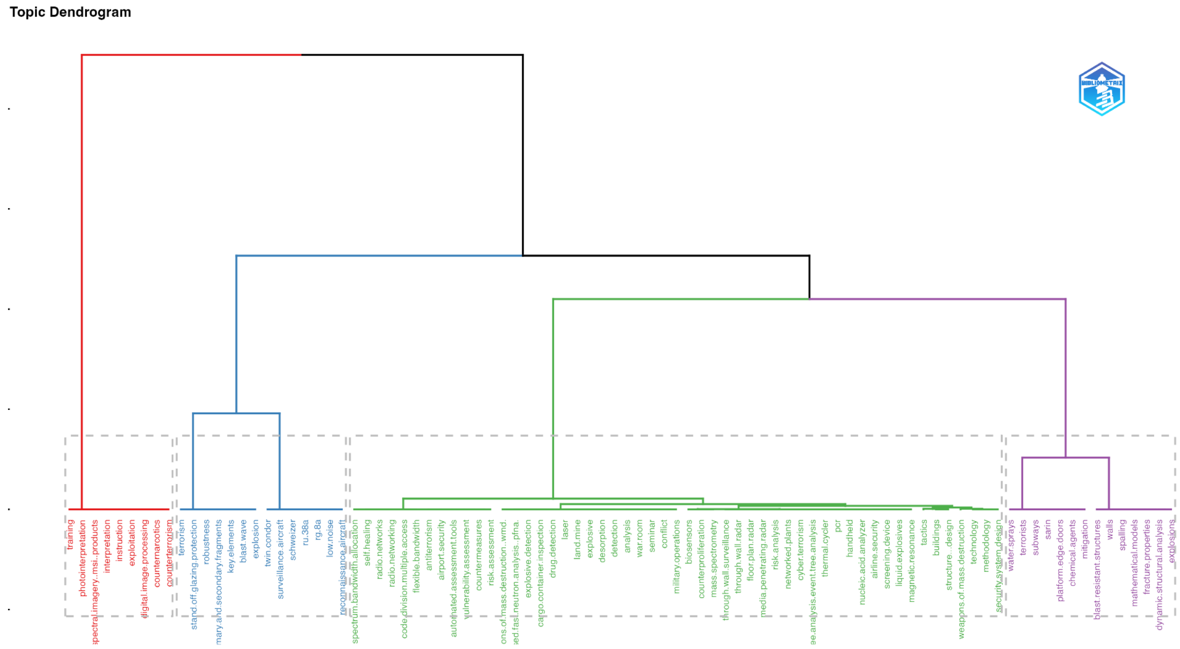


Figure 5. TRER topic dendrogram: topic keywords for 20xx-20yy.

In the next section we describe the results of a time sliced analysis of the TRER publication data. In Figure 6, cut off points are added to indicate five phases of TRER research that were identified from trend breaks in data. The first (1989-2000): dates between end of the Cold War and just before 9/11. As noted above it is characterized by a low publication output. Next phase is an accumulating phase starting in 2001 that tops out in 2006. The third phase is characterized by a declining level of output between 2007 and 2013. A fourth period of renewed increase was identified 2014–2018, with the last, ongoing decrease beginning in 2019. For each of these phases a co-word map was produced as described above that contained the most relevant noun-phrases in titles and abstracts.

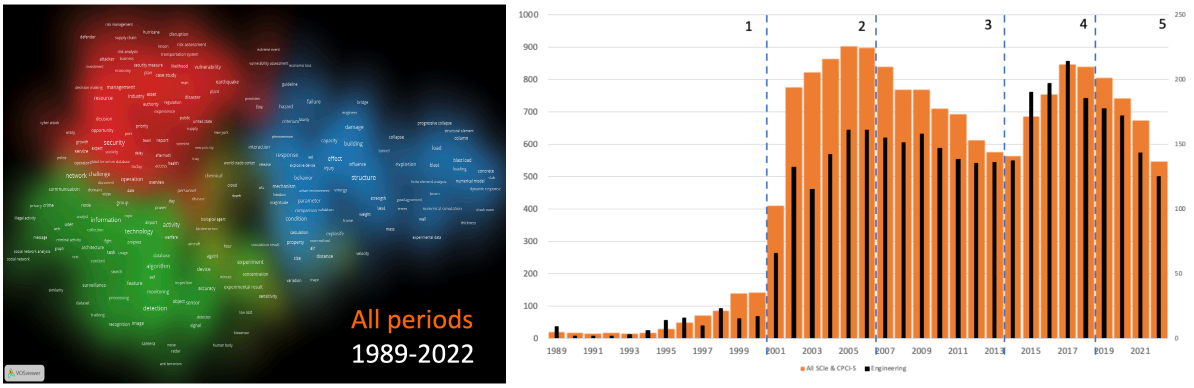


Figure 6: Cut off points used to delimit the published articles in five groups: Pre-9/11 (1), accumulating phase (2), declining phase (3), revival phase (4) and a possible second decline (5).

5.1 Phase 1: Pre 9/11 (1989-2000)

The left co-word map in Figure 7 shows the most significant terms identified in the 1989–2000 set of 121 papers. As the sample is rather low with, it is not as detailed as a map based on a larger sample could be and terms in one or few individual articles might be given a disproportionate salience even though all shown terms are are about equal size. At the center we find *terrorist threat* and very general terms (*programs, effort, decision*) as well as *end, cold war* and *soviet union*, pointing to the contemporary historical context. This is surrounded by four more distinct – albeit not very densely clustered – topics apparently connected to different terrorist threats. The rightmost blue cluster is the most distinct and interpreted as a topic on 'building explosions' which is next to a green cluster on *hacking* and *computer security,* with a yellow cluster below connected to *chemical* and *biological terrorism*. The leftmost red cluster relates to electromagnetic disturbances to civilian aircraft electronics and avionics (*em terrorism effect, airplane cabling).*

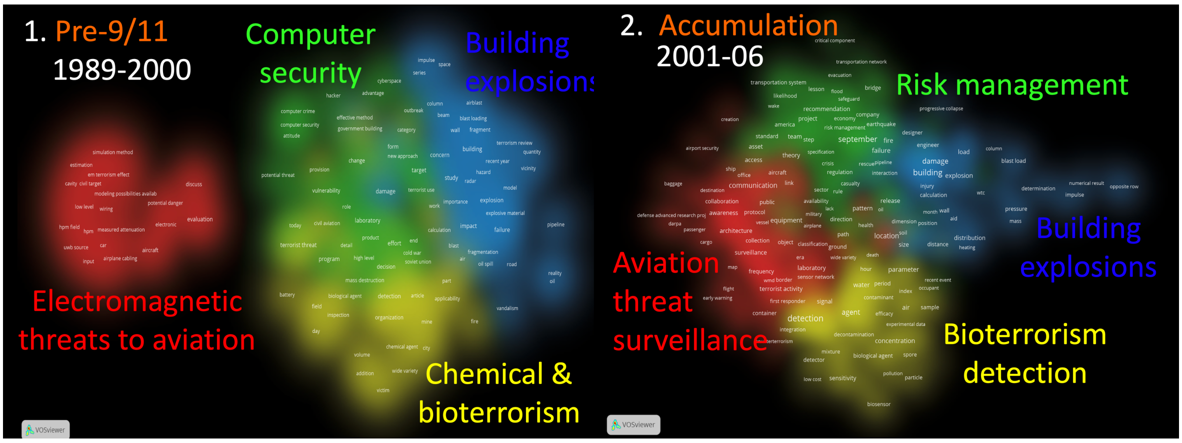


Figure 7: Co-word analysis: Phase 1: Pre 9/11 (1989-2000), n=121 noun phrases found ≥2 times. Phase 2: Accumulation (2001-2006), n=779 phrases found ≥5 times.

5.2 Phase 2: Accumulation (2001-2006)

In the second phase a lot has changed. First, the sample of papers is much bigger which results in a larger sample of noun phrases that can be related to each other and a much fuller analysis to be performed. The same colors have been chosen for the topic clusters that are found from the previous phase map. The blue cluster still relates to 'buildings explosions' and structural capacity and the yellow cluster is now focused on 'bioterrorism detection' with terms such as *biosensors, biological agent, contaminant* something which seems to be in line with a particular focus on bioterrorism in particular US government funding. The previous 'chemistry' part of the topic is not found anywhere in the graph. This could imply a lower interest in that part in this period than before.

The most striking aspect is that the two clusters on computer security and electromagnetic threats are gone and replaced by a red cluster on more general 'aviation threat detection' and a green interpreted as being a general 'risk management' cluster directed towards natural disasters and threats to infrastructure. Significant, but not surprisingly, is that all the clusters appear to relate to the 9/11 attacks, as seen by the prominent terms *september* and *wtc* of the buildings and risk management clusters, the bioterrorism cluster's *recent events* most likely indicating the post-9/11 attack in the US using ricin letters, and the aviation threat cluster through its new focus on *passenger*, *flight* and *airport security*.

5.3 Phase 3: Decline (2007-2013)

In the third phase, we have noted that the volume of published papers is gradually declining from the peak in 2006. Nonetheless there is still high frequency in publication as compared with the first phase. In total, this seven-year period yields 978 published papers in the TRER set. Content wise, this period roughly correlates with the previous ones, but in browsing the terms that are found, we could at the same time identify a specialization in terms of the phrases used, and a generalization or rather a 'reflexive theorization' in terms of what topics are discussed. This is seen in terms like *numerical simulation, game theory* and *detection algorithm* on the one hand, and on the other terms associated with human wellbeing and social issues such as *public safety, citizen, society, surveillance,* and *economy* spread across the map. The red 'detection & surveillance' cluster is still focused on prevention and detection against terrorist threats, but the phrases used are now more focused on detection terms as well as being more general without the previous passenger aviation focus.

The blue cluster still relates to buildings and structural damage, while following the trend of being more generalized than in the previous periods. The green cluster previously focused on risk management, appears to be shifting towards ‘security threat management’ handling antagonistic militant and military threats with reference to *security measure, terrorist event, potential target, defender, soldier, uav*, being placed close to country names such as *USA, Iraq* and *Afghanistan.*

The yellow ‘bioterrorism’ seems to somewhat disappear, but on the other hand the phrase bioterrorism is found prominently in the middle of the co-word map. Arguably this is due to the bioterrorism topic being subsumed within the whole terrorism discourse in research publications rather than that interest in this area has declined.

The topic dendrogram (not shown) shows many similar and indistinct cluster topics while strengthening the coword map’s focus on buildings.

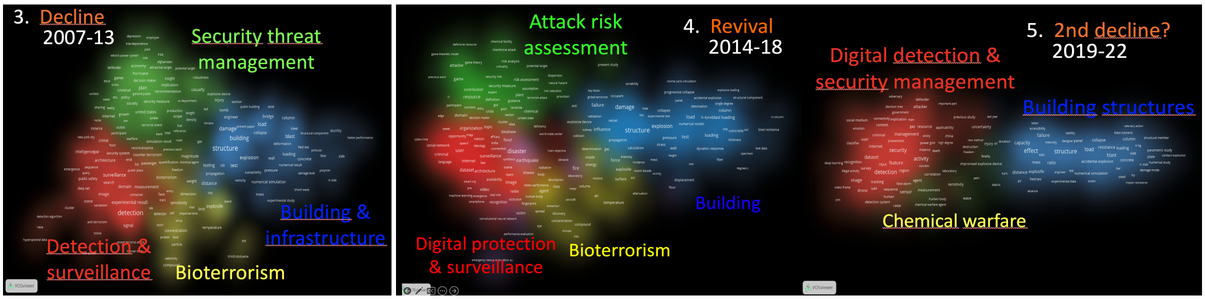


Figure 7: Co-word analysis Phase 3 Decline (2007-2013) n=942 noun phrases found ≥5 times; Phase 4 Revival 2014-2018 n=980 noun phrases found ≥5 times & Phase 5 Second Decline (2009-2022) n=725 noun phrases found ≥5 times.

5.4 Phase 4: Revival (2014–2018)

In the fourth phase, the decline we saw in the previous phase was reversed into an increase of TRER interest. In this phase we see an increase in the annual publication from less than 150 papers in 2014 to more than 200 papers in 2017. On average, this period is distinguished by a higher number of papers annually as opposed to the previous one.

Qualitatively, we note that a new cluster, here labeled ‘Digital protection & surveillance,’ has evolved from the ‘Detection and surveillance’ cluster identified before. Here, terms like *cybercrime*, *social network, dataset, convolutional neural network* and *machine learning* stand out as the most distinguished terms. The yellow ‘bioterrorism’ cluster now appears to contain more general terms and being less antagonistic. The green cluster seems now even more focused on ‘attack risk assessment’ with a remaining strong focus on antagonistic attacks although more disconnected from the previous military context. The last blue buildings cluster is still very strong.

5.5 Phase 5: Second Decline (2019–2022)

In the last phase, again, we see a new second decline in the number of papers each year. It should be noted, that at the time of extracting data from Web of Science in March 2023, all papers published in 2022 were not yet indexed and thus the staple for 2022 is most likely not complete. As noted in Figure 1, both regarding the normative line depicting the annual increase of WoS, which up until 2022 is mostly linear, and the number of publications in the TRER dataset, which shows remarkably fewer publications the year before (125 vs. 143). Therefore, some caution is called for regarding this trend.

Still, a feature that is quite clear in the last phase is that the previous bioterrorism cluster now has almost disappeared and possibly replaced by a more diffuse ‘chemical warfare’ cluster. Furthermore, the two clusters on antagonistic attack management and the detection and surveillance now seem to have grown together into a ‘digital detection and security management’ cluster, making it impossible to distinguish them in the data. Finally, the traditional building and infrastructure cluster (blue) still stands out as a distinct and substantial part of the research.

6. Conclusion

This study of the effect of terrorism in engineering research has strengthened our earlier results showing that there was an increased focus on terrorism related STEM research following the terrorist attacks on 11 September 2001. More specifically we have shown that it is possible to use distant reading methods to provide qualitative new and previously unknown knowledge about how terrorism influenced research on a global level, a beginning of a larger and more detailed history of the impact of terrorism on STEM in general and on engineering research in particular. Particularly, we have shown it is possible to discern distinct focus areas of terrorism-related engineering research before and after the end of the Cold War. The content of these research areas resonates with what we know from developments following 9/11 such as discussions among the civil engineering community (and outside) about the structural causes of the collapse of the WTC buildings and, in the USA, increased government funding towards research on bioterrorism and focus on protecting critical infrastructures. Whether the researchers behind this research actually led or responded to this development does however need further in-depth investigations. Furthermore, we also see a dynamic in that the interests change over time which points to the need to look into how researchers' attention to terrorism was stabilized or replaced over time.

More specifically, one can discern several possible case studies for further research. As noted above, the TRER dataset is interesting in the way that older research is not more frequently cited than newer research. Of course, this is mainly due to the number of published papers during the first period (1989–2000) is much lower than the subsequent two periods. Even so this is quite remarkable and requires a specific study. We propose to investigate this by following the few papers that are published before 2001 that are cited in the post 2001 literature. There are four papers in the set that match these criteria (as well as a fifth that is cited within the pre-2001 time frame). Together, these papers are cited 16 times by thirteen other papers. An interesting issue in connection to this would be to study how many of the post-2001 articles have references to *September*, *World Trade Center,* or *WTC*? One finding was that there is a 'semantic drift' over time, that also can be attributed to 'incorporation by obliteration', where terms that earlier were found as buzzwords (e.g. 'terrorism', proper), is developed into more specific notions of the phenomena that is attributed. A distinct feature is that new and alternate terms than terrorism were formulated after 2001 for new terms for terrorism-related or terrorism-like activities such as 'homeland security', 'radicalization' or 'violence-affirming extremism' [40]. A specific study focusing on these terms in the publications, as well as on acknowledgments of research funded from the Department of Homeland security should reveal whether they replace or complement the earlier research. From 2003 we can find the new term 'homeland security research' in our data.

Finally, we have shown from a methodological perspective that exploratory bibliometrics and algorithmic historiography can complement traditional qualitative historical research methods, suggesting new ways of gaining insights into historical phenomena by assisting in identifying relevant connections and relationships not immediately apparent through traditional qualitative analysis and by generating hypotheses for further exploration. In this way showing the potential of scientometric methods to historical research, we see this methodological introduction as a central contribution towards expanding the toolbox not just of digital historians but of digital humanists overall.

7. Acknowledgements

This study is part of the research project *Things for living with terror: a global history of the materialities of urban terror and security* (P18-0912:1) funded by the Swedish Foundation for Humanities and Social Sciences (RJ) and the University of Gothenburg.

8. References

[1] Hounshell, D. A. (2001). 'Rethinking the Cold War', *Social Studies of Science*, *31*(2).

[2] Davis, M. (2001). 'The flames of New York'. *New Left Review* 12 (November-December).

[3] Reppy, J. (2008). 'A biomedical military–industrial complex?'. *Technovation*, *28*(12).

[4] Moreno, J. D. (2012). *Mind wars*. 2nd rev. ed. Bellevue Literary Press.

[5] Fridlund, M (2011). 'Bollards, buckets and bombs', *History and technology*, *27*(4).

[6] Fridlund, M. & Nelhans, G. (2011). 'Terrorens ingeniørkunst', in M. Fenger-Grøndahl, ed., *11. september: Verdens tilstand ti år efter.* Aarhus Universitetsforlag.

[7] Fridlund, M. & Nelhans, G. (2011). 'Science and 'the 9/11-effect'', *Science Progress* Center for American Progress. *www.scienceprogress.org /2011/09/science-and-the-the-911-effect/.*

[8] Fridlund, M. & Nelhans, G. (2011). 'Naturvetare i kriget mot terrorismen', *Tvärsnitt 33*(3-4).

[9] Moretti, F. (2000), 'Conjectures on world literature', *New Left Review* N.S., *1*(1).

[10] Moretti, F. (2013), *Distant reading*, London: Verso.

[11] Price (1961). *Science since Babylon*. Yale University Press.

[12] Price, D. J. d. S. (1963). *Little science, big science*. New York: Columbia Univ. Press.

[13] Price D. J. d. S. (1967). 'Research on research'. In D. L. Arm, ed., *Journeys in Science.* Uni. of NM Press.

[14] Nelhans, G. (2013). *Citeringens praktiker.* Ph. D. thesis, University of Gothenburg, Gothenburg.

[15] Woolgar, S. (1991). 'Beyond the citation debate'. *Science and Public Policy 18*(5).

[16] Merton, R. K. (1973). *The sociology of science.* The University of Chicago Press.

[17] Merton, R. K. (1988). 'The Matthew effect in science, II'. *Isis, 79.*

[18] Cole, S., J. R. Cole. (1967). 'Scientific output and recognition'. *Am. Sociological Review 32*(3).

[19] Edge, D. (1979). 'Quantitative measures of communication in science'. *History of Science 17.*

[20] Gilbert, G. N. (1977). 'Referencing as persuasion'. *Social Studies of Science 7*(1).

[21] Gilbert, G. N., & S. Woolgar, S. (1974). 'The quantitative study of science'. *Science Studies 4*(3).

[22] MacRoberts, M. H., & MacRoberts, B. R. (1989). 'Problems of citation analysis'. *Journal of the American Society for Information Science 40*(5).

[23] Latour, B. (1976). 'Including citations counting in the system of actions of scientific papers'. *1st annual meeting of the Society for Social Studies of Science,* 4-6 *Nov.* 1976. Cornell University.

[24] Latour, B. (1987). *Science in action.* Harvard University Press.

[25] Callon et al. (1986). *Mapping the dynamics of science and technology*. Macmillan.

[26] van Heur, B. et al. (2012). 'Turning to ontology in STS?' *Social Studies of Science, 43*(3).

[27] Latour, B., et al. (2012). ''The whole is always smaller than its parts' – a digital test of Gabriel Tardes' monads'. *The British Journal of Sociology, 63*(4).

[28] Basson, I., Simard, M. A., Ouangré, Z. A., Sugimoto, C. R., & Larivière, V. (2022). 'The effect of data sources on the measurement of open access'. *PLoS one*, *17*(3),

[29] Ángeles Oviedo-García, M. (2021). 'Journal citation reports and the definition of a predatory journal', *Research Evaluation*, *30*(3)

[30] Garfield, E. (1997). 'Editors are justified in asking authors to cite equivalent references from the same journal'. *BMJ*, *314*(7096).

[31] Garfield, E. (1971). 'Citation indexing, historio-bibliography, and the sociology of science'. In E. Garfield, ed., *Essays of an Information Scientist Vol. 1*. Philadelphia.

[32] Garfield, E. (2009). 'From the science of science to Scientometrics: Visualizing the history of science with HistCite software'. *Journal of Informetrics, 3*(3).

[33] Garfield, E., Pudovkin, A. I., & Istomin, V. S. (2003). 'Why do we need algorithmic historiography?' *Journal of the American Society for Information Science and Technology, 54*(5)

[34] Garfield E., Sher I. H. & Torpie R. J., (1964). *The use of citation data in writing the history of science,* Institute of Scientific Information.

[35] Merton, R. K. (1991) [1965]. *On the shoulders of giants.* University of Chicago Press.

[36] Leydesdorff, L. (2010). 'Eugene Garfield and algorithmic historiography'. *Annals of Library and Information Studies*, *57*(3)

[37] Van Eck, N. J., & Waltman, L. (2010). 'Software survey: VOSviewer, a computer program for bibliometric mapping'. *Scientometrics, 84*(2).

[38] Van Eck, N. J., & Waltman, L. (2011). 'Text mining and visualization using VOSviewer'. *ISSI newsletter, 7*(3).

[39] Aria M., & Cuccurullo C. (2017). 'Bibliometrix'. *Journal of Informetrics*.

[40] Andersson, D. E. (2018). 'Från terrorism till våldsbejakande extremism'. In M. Arvidsson, L. Halldenius, & L. Sturfelt, eds., *Mänskliga rättigheter i samhället*. Bokbox förlag.

Automated Coding of Historical Danish Cause of Death Data Using String Similarity

Louise Ludvigsen[1], Mads Perner[1,2], Bjørn-Richard Pedersen[3], Rafael Nozal Cañadas[4], Anders Sildnes[4], Nikita Shvetsov[4], Trygve Andersen[3], Lars Ailo Bongo[4] and Hilde Leikny Sommerseth[3]

[1]*Saxo-Institute, Department of History, University of Copenhagen*

[2]*Section for Data Dissemination, Danish National Archives, Odense, Denmark*

[3]*Department of Archaeology, History, Religious Studies and Theology, UiT The Arctic University of Norway*

[4]*Department of Computer Science, UiT The Arctic University of Norway*

Abstract

The study of causes of death has been central to some of the most influential studies of the modern mortality decline in the nineteenth and twentieth centuries. The digitization of individual-level cause-of-death data has been game-changing, however, the data presents a major challenge: how do we code the thousands of unique strings for analysis in an efficient way? This paper aims to see how far we can get with automated coding based on string similarity. We do this by applying a Jaro Winkler string similarity algorithm in Python (pyjarowinkler) that codes our cause of death data from the Copenhagen Burial Register 1861-1911 to DK1875, a contemporary coding and classification system from nineteenth century Denmark. We then compare the performance of the algorithm to that of a manual (historian) coder in three different ways: at the level of each unique cause-of-death string, at the level of each cause-of-death group and for the overall cause-of-death pattern for all burials in Copenhagen 1861-1911. Our results show that a minimum-effort algorithm coded approximately half of the causes of death correctly compared to the manually coded dataset. This means that the method applied here is not accurate enough to use for actual data analysis of mortality patterns, as it is not possible to examine individual causes within larger causal groups. However, the results are promising for different uses of the method as a help for the manual coder. A way forward could be to use cut-off points of the Jaro-Winkler scores, coding only those causes where the string similarity match is relatively certain or use the automated method to catch most of the initial cases of a certain disease with a very set phrasing, such as cancer. In both cases, the remainder of the unique cause of death strings could then be coded by a manual coder.

Keywords

historical causes of death, string similarity, automated coding, mortality, individual-level data

DHNB2023 | Sustainability: Environment - Community - Data. The 7th Digital Humanities in the Nordic and Baltic Countries Conference. Oslo – Stavanger – Bergen, Norway. March 8–10, 2023.

louise.ludvigsen@hum.ku.dk (L. Ludvigsen)

0000-0002-8708-4416 (L. Ludvigsen); 0000-0003-3890-207X (M. Perner); 0009-0000-2363-0791 (B. Pedersen); 0000-0001-6492-0218 (R. N. Cañadas); 0009-0003-0141-6112 (A. Sildnes); 0000-0002-8472-3702 (N. Shvetsov); 0009-0008-0893-367X (T. Andersen); 0000-0002-7544-2482 (L. A. Bongo); 0000-0001-7070-8184 (H. L. Sommerseth)

DHNB Publications, DHNB2023 Conference Proceedings, https://journals.uio.no/dhnbpub/issue/view/875

1. Introduction

Historical causes of death are key to our understanding of one of the most revolutionary developments in population history: the great mortality decline of the nineteenth and twentieth centuries. Both Omran's theory of "The Epidemiologic Transition" [1] and McKeown's "The Modern Rise of Populations"[2] rely heavily on historical cause of death data in their attempts to describe the processes driving the mortality decline. One point of criticism has been their use of relatively sparse source material, consisting mainly of official aggregated statistics, published in the late nineteenth century and early twentieth century when these historical processes were still taking place [3, 4, 5, 6]. Contemporary statistics do indeed allow us to track mortality from certain diseases, but we are limited by the categories already defined by the physicians and statisticians who compiled the statistics at the time, according to their medical paradigms and knowledge. While some cause of death categories may be relatively straightforward to work with, such as "Mæslinger" ["Measles"], others hide a wide variety of diseases, such as "Andre sygdomme i ydre dele" ["Diseases in the outer parts of the body"][1]. The problem is especially acute when trying to study mortality over long periods of time or in different places, as parallel with actual changes in the disease landscape, the coding groups changed as the medical field developed. These difficulties are central in the discussion on how to work with historical causes of death, and what can be gained from doing so [8, 9, 10, 11, 12, 13].

Following years of work by archives, professional transcribers and the engaging community of citizen science and volunteers on digitizing historical sources, historical scholars now have access to unprecedented datasets and materials in both size and detail [2]. It has been a game-changer for long-term mortality studies that individual-level causes of death from historical sources such as parish registers, death certificates and burial records are now being digitized. With individual-level causes of death from the original handwritten records, we now have the opportunity for a much more in-depth analysis of the changes in disease patterns and in the factors that influence mortality risk in a given population and across generations.

In this paper, we make use of the Copenhagen Burial Register, which has been digitized and transcribed for the period 1861-1911 by the Copenhagen City Archives [15]. It contains the handwritten record of over 300,000 individual burials and more than 10,000 unique causes of death. This means that for the first time, it is possible to work with Danish individual-level causes of death from the nineteenth century at a large scale. However, the data presents a major challenge: how do we code the thousands of unique strings for analysis in an efficient way? Historians have traditionally preferred manual coding and consider this a gold standard. It is, however, a very time-consuming process, and since it also requires a considerable level of domain expertise, the method does not scale well and is unsustainable in the long run. One possible solution is to use machine learning algorithms, as has already been shown by several other projects on both historical and modern causes of death [16, 17, 18, 19]. However, machine learning is typically based on training data, which still relies on a time-consuming manual coding process by a person with large domain expertise. In addition, the earlier research

[1] The examples are taken from the Danish DK1875 system, used in the Danish aggregated cause of death statistics from 1876-1930. See appendix A and reference [7].

[2] An example of this is the SHiP collaboration working on individual-level causes of death in multiple European countries [14].

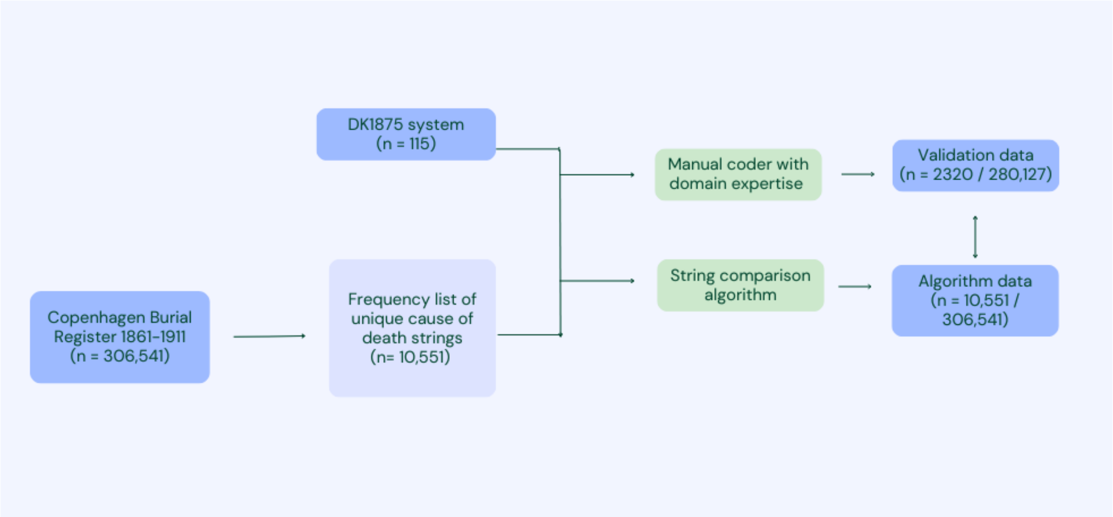

DK1875 system
(n = 115)
Copenhagen Burial Register 1861-1911 (n = 306,541)
Frequency list of unique cause of death strings (n= 10,551)
Manual coder with domain expertise
String comparison algorithm
Validation data (n = 2320 / 280,127)
Algorithm data (n = 10,551 / 306,541)

Figure 1: Diagram of the sources used in this paper.

focuses heavily on automated coding to the modern ICD-10 system, which does not give us any information on how an automated coding algorithm would work within the context of a historical coding and classification system.

This paper aims to contribute to the discussion on automated coding by exploring how accurately we can code historical causes of death to a historical classification system with the relatively simple technique of string similarity matching. Using a historical classification system is essential for string similarity techniques, as a modern system such as the ICD10 of ICD10h [14] will either be in a different language or have very different phrasings compared to the contemporary causes of death on the death certificates. We therefore apply a set of automated string similarity algorithms to code our cause of death data to "DK1875", a coding and classification system developed for use by nineteenth century physicians in Denmark (see appendix 1 for the full DK1875 system). This is done without any previous training of the algorithm, and thus without any use of training data. Since the causes of death in the Copenhagen Burial Register have already been manually coded to DK1875, we compare the performance of the algorithm to this validation dataset, constructed by a manual (historian) coder with extensive domain expertise (Figure 1).

The paper is based on work done in a 48-hour hackathon at UiT The Arctic University of Norway in May 2022, in which all authors participated. At the hackathon, two small teams were competing to solve the following challenge: code the 10,000+ unique causes of death in the Copenhagen Burial Register to the correct category in the DK1875 system. In the hackathon and in this paper, we have defined 'correct' coding as one that matches the coding done by hand in the validation data. To tackle the issue, both teams were given 1) a frequency list of all unique causes of death strings in the burial data and 2) a list of the 115 categories of the DK1875 system. Both teams chose string similarity matching to assign a code to the causes of death, but their methods differed slightly. One team did only little data cleaning, relying instead on a range of different string similarity methods to find the best fit for each cause of death based on confidence scores. The second team did more extensive data cleaning and relied upon a single-string similarity measure. In this paper, we rely on the latter approach to improve automatic coding and study which possibilities and limitations it creates, as it seems to provide the most consistent results[3].

2. The data

2.1. The Copenhagen Burial Register

In 1861 the Copenhagen municipality introduced the Copenhagen Burial Register to centralize the administration of the growing number of burials in the city caused by rapid growth in population. Previously the burials had been recorded in individual registers for each cemetery, but with the introduction of the Copenhagen Burial Register, all burials within the city (except for the burials at military, Roman-Catholic and Jewish cemeteries) were to be recorded in one volume. From 1887 the registration rules were changed, and from then on, the Copenhagen Burial Register contains all deaths that happened in the city, regardless of where the burial took place [20]. The Burial Register is remarkably consistent; no volumes are missing, and they contain close to the same information for each burial throughout the period. The information includes name, age, occupation, date of death, cause of death, burial date and place, address at death and where the deceased's body was kept prior to burial [21]. Earlier studies have compared the information to that found in the equivalent death certificates and concluded that while the information is mostly the same, the Copenhagen Burial Register provides more complete personal information [22].

The individual-level burial records contained in the register have first been digitized, and then transcribed by volunteers at the Copenhagen City Archives, resulting in scanned facsimiles and a machine-readable version with a link between the two versions of the same burial [20]. For the period from 1861 to 1911, the database contains a total of 306,541 burials and 10,551 unique causes of death. When transcribing the causes of death, volunteers were instructed to do it "…i den rækkefølge de står og skrives kildetro, på nær forkortelser, variationer i stavemåder og åbenlyse stavefejl." ["…in the order they appear and true to the source, except for abbreviations, variations in spelling and obvious mistakes in spelling"] [23]. The transcriber had the option to either choose from a drop-down list of previously used cause-of-death values in the source or to write a free-text alternative themselves if the drop-down list did not contain the cause of death they were transcribing. Finally, a quality control procedure handled by so-called 'super-users' has corrected errors and ensured that the transcription always corresponds to the handwritten word or string on the record. All these initiatives contributed to standardizing and cleaning the causes of death as a part of the transcription process while interfering as little as possible with the content of the source. As a result, the transcribed causes of death from the Copenhagen Burial Register include different spellings for the same word into one string, such as "ukendt (ubekendt, ubekjendt)" ["unknown"], while separating synonyms for

[3] Python scripts and data used for the analysis in this paper are available in the following GitHub repository link: https://github.com/louiseludvigsen/Automated-Coding-of-Historical-Danish-Cause-of-Death-Data.

Table 1
Example of the frequency list of unique cause of death strings

ID	Tidy cod	Freq	Acc.sum	Acc.perc
1	morbus cordis (mb. cordis, mb. cord.)	15,100	15,100	3.44
2	dødfødt	13,290	28,390	6.47
3	pneumonia (pneumoni)	11,813	40,203	9.16
4	..	..	..	..

the same illness, for example: “Engelsk syge” [Danish term for rachitis] and “Rachitis”. Further, if a particular cause of death has a descriptive attachment, it is listed as a separate string, for example: “Nephritis” and “Nephritis chronica”. Overall, the transcription setup of the archives has lessened the need for data cleaning, but some issues remain for doing a string comparison, as will be explained in the methods section.

2.2. The frequency list of unique cause of death strings

A frequency list of all the unique cause-of-death strings that were used for the 306,541 burials in the Copenhagen Burial Register 1861-1911 was created by Ludvigsen, without considering whether the cause-of-death string was used as first, second, third etc. cause of death. The list contains 10,551 rows (one for each unique cause of death string) with the following information:

- tidy cod: the unique cause of death string
- Freq: frequency of use as the primary cause of death
- Perc: The percentage of burials with this string as the primary cause of death
- Acc. Freq: accumulated frequency of use as the primary cause of death
- Acc. Perc: accumulated percentage of burials with these strings as the primary cause of death.

This list was used unconnectedly for the automated string comparison and for Ludvigsen to create the validation data.

2.3. The validation data

The validation data was created based on the frequency list described above, by Louise Ludvigsen, a historian with large domain expertise in the registration and coding of historical causes of death in nineteenth- and twentieth-century Denmark. Ludvigsen manually coded 2,320 of the 10,551 unique cause of death strings in the list, by order of frequency starting with the most frequent, using about 6 months to do so. The 2,320 manually coded cause of death strings are used as the primary cause of death for 280,127 of the 306,541 burials (91.38%).

Each of the 2,320 unique cause of death strings was coded to three different coding systems in two separate steps. First, the unique cause of death was coded to the ICD10h coding scheme, constructed by the SHiP network, adapting an earlier coding scheme developed by the Cambridge Group for the History of Population and Social Structure [14]. The ICD10h is largely based on the ICD10 system designed by the World Health Organization (version 2016), but it contains two additional digits at the end of each category to allow for more historical nuance (e.g. A16.904) [14, p. 68].By default, the causes were assigned ICD10 codes as well since they are essentially identical to ICD10h save the suffix. Secondly, the same unique cause of death string was coded to the contemporary Danish system DK1875, used in the official Danish cause of death statistics from 1876-1930 [24, p.188]. As mentioned previously, we only make use of the DK1875 classification system in this paper. Using a modern system such as the ICD10 of ICD10h will not work for our use of a string similarity technique, as they will either be in a different language or have very different phrasings compared to the contemporary causes of death on the death certificates.

2.4. The DK1875 system

The DK1875 system was introduced in the official Danish cause of death statistics on the 1st of January 1875 and consists of 115 categories, placed within nine main groups (see appendix A). The 115 categories describe either a singular disease or groups of diseases, such as "Kighoste" ["Whooping cough"] and "Andre Farsoter" ["Other epidemic diseases"]. Each category has a number, a Latin description, and a Danish description (see appendix A). The system was inspired by the contemporary English and Swedish systems, and particularly by the principles of the English statistician Farr [24, p.165]. It is multifaceted as it works simultaneously as a nomenclature, a coding scheme, and a classification system. It serves as a nomenclature, as the physicians were instructed to use the 115 categories when filling out the death certificates [25], and as a coding and classification scheme when used in the official cause of death statistics and other statistical analysis, either as the 115 categories or added together into larger groups for analysis. The DK1875 system received only very minor updates before it was ultimately replaced by a common Nordic system in 1931 [24, p. 193].

3. Methods

Based on our experiences from the hackathon at UiT The Arctic University of Norway, we decided to apply a single method for the automated coding in this paper. We have employed the Jaro-Winkler string similarity measure, which has been shown to be effective in matching strings with typographical errors and inconsistencies. The Jaro-Winkler measure is based on the Jaro distance, which calculates the similarity between two strings based on the number of matching characters and the number of transpositions required to make the strings identical. The Winkler modification adds a scaling factor based on the length of the common prefix of the strings, which accounts for the likelihood that two similar strings have a common prefix. Due to these features, we expect Jaro-Winkler to perform better than other string similarity measures, such as Levenshtein distance and cosine similarity which only captures edit distance, since Jaro-Winkler is more robust to differences in string length and minor typographical errors.

3.1. Data cleaning and string comparison

There were three parts to the data cleaning in the frequency list and the DK1875 list: abbreviations, information in parentheses and translating the Nordic letters æ, ø and å. In the frequency list, we removed the accents and special characters from the unique cause of death strings, as well as transliterating the Nordic letters æ, ø and å to ae, o and a, respectively. In addition, we moved the contents of the parentheses in the unique cause of death strings to a new column since they would otherwise disturb the string-matching algorithm. The information in the parentheses either contained an abbreviation, such as "tuberculosis pulmonum (tub. pulm.).", or the Latin phrase for a cause written in Danish (or vice versa). In the DK1875 coding scheme, cleaning was done for both the Latin and Danish terms for each code. First, abbreviations were written out, e.g. "bronchit." became "bronchitis". Secondly, we shortened categories with lengthy names when we suspected it might improve the matching. And finally, we separated categories when they contained several specific causes of death, such as "Cholerine & catarrhus intestinales," ["Domestic cholera and acute diarrhoea"], which were separated into "Cholerine" and "catarrhus intestinalis".

After cleaning the data, we used the 'pyjarowinkler' package [26] in Python to calculate the Jaro-Winkler distances, which returns a score between 0 and 1, with 1 being the highest similarity between the two strings. For each unique cause of death string, we computed Jaro-Winkler distances between the string and both the Danish and Latin labels of each of the 115 DK1875 categories. The string was then assigned the DK1875 code with the highest density score. No minimum score was set for assigning a code, and thus all 10,551 unique cause of death strings were assigned a code.

3.2. Testing the accuracy

To test the accuracy of our approach and the consequences it may have for studies on historical mortality, we have compared the codes assigned by the automated string comparison method to the validation data in three different ways: at the level of each unique cause-of-death string, at the level of each Heiberg group and for the overall cause-of-death pattern for all burials in Copenhagen 1861-1911.

To test the accuracy of our approach at the level of each unique cause of death string, we compared the DK1875 code assigned by the string comparison algorithm directly to the manually assigned DK1875 codes in the validation dataset which is our "ground truth" data. If the two codes were the same, we considered the code assigned by the algorithm correct. If they were different from each other, we considered the code assigned by the algorithm wrong. As not all 10,551 strings have been coded by hand, the comparison was based on the 2320 strings that were.

However, in practice, the 115 subgroups are rarely used directly for analysis. Since there are too many of them, and some describe groups of diseases while others describe individual diseases, it is more convenient to classify them into larger groups. In this paper, we have used the classification system presented by the Danish physician Povl Heiberg in his study of mortality among 15–74-year-olds in Denmark in the 1890s and early 1900s (from here on referred to as the Heiberg groups)[27]. Heiberg himself assigned each of the 115 DK1875 subgroups to one of twelve groups in his classification scheme (see appendix B). We are interested in two measures of coding accuracy here: 1) the proportion of the codes in each Heiberg group that have been coded accurately according to the validation data. This is to see if certain disease groups are more difficult to code for the automatic method. 2) the proportion of the individual-level causes of death that have been assigned to the correct Heiberg group, according to the validation data, even though the DK1875 might be inaccurate. This will help us measure when the algorithm has not coded to the same code as the validation data, but to an adjacent category that may be very similar. Using the Heiberg groups allows us to look closer at what effects the accuracy will have for studies on historical mortality, which often makes use of larger groups like these for analysis.

Finally, we have examined how the automated coding algorithm performs when analyzing the cause of death patterns for the burials of people aged 0-80 in Copenhagen 1861-1911. For this analysis only, the burials of people over age 80 have been excluded, and the Heiberg groups “suicides”, “accidents” and “genitourinary diseases” were put into the Heiberg group “others”.

4. Results

4.1. Coding of the unique cause of death strings

Of the 2,320 coded causes of death in the validation data, our algorithm managed to code 1,075 (46.3%) strings correctly (i.e. in the same way as the manual coder). If we look at the number of burials coded correctly, rather than the unique causes of death, the number is somewhat higher, as the algorithm was able to code 176,681 of the 280,127 burials (63%) correctly compared to the validation data. There are more correct codes for the burials than there are at the level of each unique cause of death since the automated approach overall tends to be slightly more accurate in coding the causes that most frequently appear in the burials. Ordered by frequency, 283 of the 500 most frequent causes were coded accurately (56%), while the number was 667 out of the remaining 1,529 causes of death (44%). This probably reflects the fact that the more frequent causes tend to be shorter and simpler, while the less frequent causes are often more elaborate in detail, and thus differ more from the categories in string similarity.

The 100 most common unique causes of death strings (roughly 1% of all the total unique cause of death strings) account for more than 78% of all the burials in Copenhagen from 1861-1911. Among these 100, the automated string comparison was able to assign a correct code for 65 out of 100 of the unique cause of death strings compared to the validation data (Figure 2). The accurately coded cause of death strings tends to be relatively short and have a very high overlap with the wording in the DK1875 descriptions. They are also mostly distinctive causes of death in the sense that there is only one category in the DK1875 system that will match, and not several for the algorithm to choose between. Smallpox, for instance, has a single code that all cases are assigned to. Tuberculosis, on the other hand, has several codes in the DK1875 system, accounting for different variations of the disease: “29: Akut miliærtuberkulose” [acute miliary tuberculosis], “30: lungevindsot” [pulmonary phthisis] and “31: tuberkulose i andre organer” [tuberculosis in other organs]. For a simple algorithm like the string similarity method applied here, it is very difficult to distinguish between these.

For the top 10 most frequent unique cause-of-death strings, the string comparison algorithm

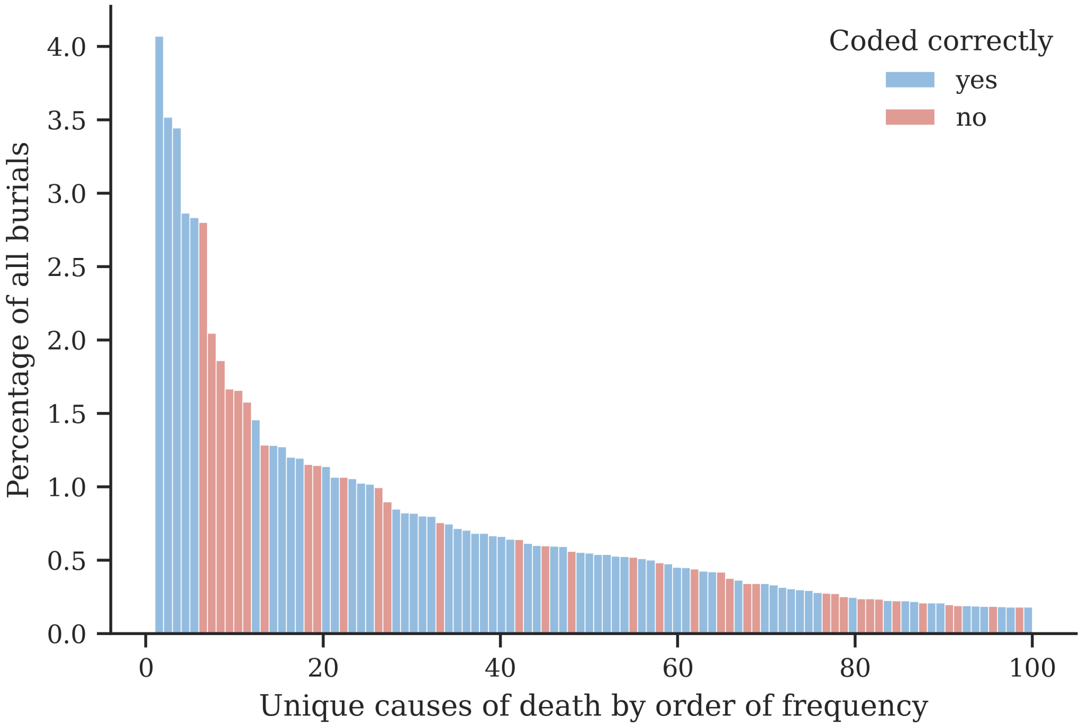


Figure 2: Automated string comparison coding compared to validation dataset, for the 100 most frequent causes of death in the Copenhagen Burials 1861-1911.

assigns a wrong code in half of the cases (Figure 2). Most of these are relatively easy to code manually, but the algorithm gets them wrong because the strings are completely different in the frequency list and the DK1875 system. This is the case for the strings “kramper” ["cramps"], “ukendt” ["unknown"], and “pludselig død” ["sudden death"] in the frequency list. “Kramper” [“cramps”] is a Danish term for “convulsioner” ["convulsions"], which has an exact match in the DK1875 categories. Similarly, the proper code for “ukendt” ["unknown”] would be no. 113, “Uangivet eller slet specificeret dødsårsag” ["Unnoted or poorly specified cause of death"]. In this case, the language of the DK1875 system is not only too different, but also too elaborate for a string similarity method to work. In other cases, the automated method fails because the causes in question fit into several categories, once again caused by the DK1875 system being too elaborate. This is the case for “tuberculosis pulmonum” [“pulmonary tuberculosis”], “meningitis” and “bronchitis”. Tuberculosis has three different categories in the DK1875 system as mentioned previously, while meningitis has two different categories, “61, hjernebetændelse” [“cerebral meningitis”] and “69, Rygmarvsbetændelse” [“spinal meningitis”], which both contain the word meningitis in the Latin phrasing. Bronchitis has three different categories, which all contain the word bronchitis, but distinguish between acute, chronic, and capillary bronchitis: “76: Brystkatarrh, akut bronchitis” ["Chest catarrh, acute bronchitis"], “77: Kapillær Bronchitis og katarrhalsk Lungebetændelse” ["Capillary Bronchitis and catharrhal pneumonia"], “78, Chronisk

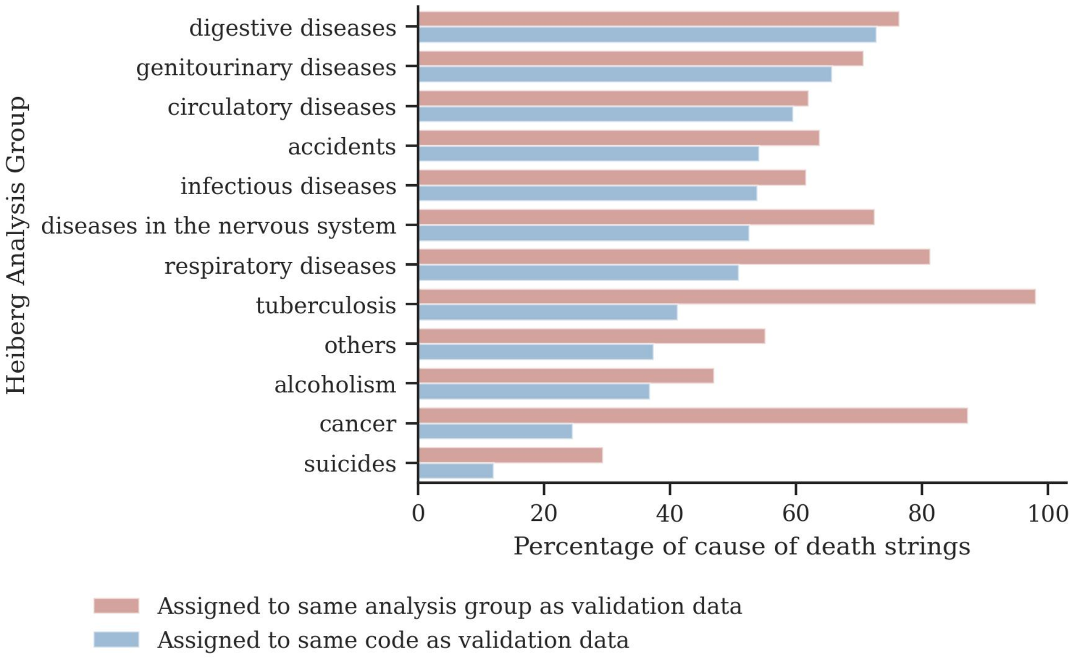


Figure 3: Measures of accuracy for each group in the Heiberg scheme.

Bronchitis” ["Chronic bronchitis"]. All three examples confirm that it is very difficult for a simple string similarity method to distinguish between multiple categories for variations of the same disease.

4.2. Coding according to Heiberg groups

As mentioned in the methods section, we are not only interested in how the string comparison method performs for each unique cause of death string, but also how it performs when looking at larger groups for analysis, such as the Heiberg groups. To see if certain disease groups are more difficult to code for the automatic method, we look at the proportion of the codes in each Heiberg group that have been coded accurately according to the validation data. For this code-specific accuracy, that is the proportion of the cause of death strings that the automated method has coded to the same code as the validation data, there is quite a bit of variation between the disease groups (Figure 3). While the automated method performs relatively well for digestive, circulatory, and genitourinary diseases, with an accuracy of c. 60-70%, it performs poorly for suicides, tuberculosis, and cancer. It is not surprising that suicides are difficult for the algorithm to code since the burial records most often contain a description of the manner of the suicide, rather than the word itself.

The pattern is different if we look at accuracy according to Heiberg’s groups, meaning causes of death where the algorithm has hit the same Heiberg group as the validation data, regardless of the specific DK1875 code. For tuberculosis, the accuracy is almost 100% at the group level, compared to ‘just’ 40% at the code level. Similarly, cancer diseases, for which the algorithm performs very poorly at the code level, are coded to the right Heiberg group almost 90% of the time (Figure 3). This striking difference in accuracy between the DK1875 codes and the Heiberg groups reflects that while the codes are wrong according to the validation data, they are coded to an adjacent, and probably very similar, code, and thus end up in the same classification group. For instance, the most common code-level error for cancer diseases is when the algorithm has placed a cause of death in codes like "Breast cancer", "Stomach cancer" or "Cervical cancer" when the appropriate is a more general one: "Cancer in other organs". This suggests that overall, the automated method is actually very accurate at capturing causes of death related to cancer, but due to the detail of the coding scheme, it is often mistaken in the individual code. The reason why the code- and group-specific accuracy for tuberculosis varies so much is the same: since there are a handful of tuberculosis-related codes in the DK1875 system, the algorithm often picks the wrong one, because it chooses the one with the shortest edit distance, but it rarely picks one that is unrelated to tuberculosis.

For the categories where the code- and group-level accuracy is similar, like digestive and circulatory diseases, this reflects that when the algorithm has chosen the wrong code, it is usually very far off, and not just another code within the same Heiberg group. For example, within the Heiberg group of digestive diseases several of the unique cause-of-death strings that in the validation data have been coded to intestinal ruptures, have been coded to both cancers and bronchitis with the string comparison method. Likewise, the most common faulty coding of heart disease is to encephalitis.

4.3. Coding according to the cause of death patterns

Overall, the cause of death pattern based on the codes assigned by the string comparison method looks fairly similar to the pattern based on the manual codes. This is very remarkable, considering that the string comparison method only assigned a correct code for 63% of the burials compared to the validation data (Figure 4).

However, on closer inspection, there are several clear differences between the patterns produced by the two methods. The cause of death pattern based on the string comparison method has more burials coded to alcoholism and cancer than the manual method. This is probably because the cause of death strings for alcoholism and cancer is often written in a multitude of ways (all containing the word alcoholism or cancer), making most of them appear quite far down the frequency list, and thus less likely to be coded by the manual coder. In addition, the automated method also has more burials coded to respiratory diseases, digestive diseases, infectious diseases, and diseases in the nervous system, but fewer burials coded to other diseases. These differences are particularly clear in the first and last years of the period. This might be because the classification system used is from 1876, which means the years before are likely to have slightly different phrasings for the causes of death, which could also explain the final years, where new knowledge and diagnoses might be hard to match with the old classification system. In combination, these differences mean that when we analyse the cause of death patterns and their development over time, the results are quite different for the

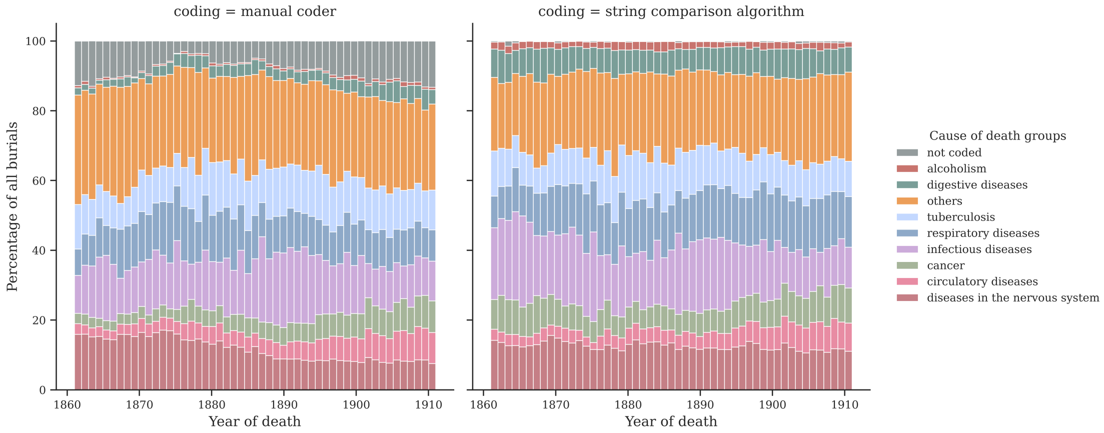


Figure 4: Cause of death pattern for the ages 0-80 in Copenhagen 1861-1911.

two methods, with the pattern from the string comparison method seeming much more stable throughout the period.

Some of the differences between the two cause of death patterns may be explained by the fact that the string comparison algorithm has coded all burials, whereas the manual coder has only coded 91.38% of the burials. However, the results from the analysis of the Heiberg groups and the unique causes of death has shown that only 46.3% of the unique causes of death are coded correctly, and that some causal groups are more affected than others, which suggest that the differences are more likely to be because of wrongly assigned codes.

5. Concluding remarks

The aim of this paper was to explore how accurately we can code historical causes of death to a historical classification system with the relatively simple technique of string similarity matching. We found that our algorithm managed to code 46.3% of the unique cause of death strings and 176,681 (63%) of the burials correctly compared to the validation data. Examining the coding of each unique cause of death string, the automated method performs relatively well for digestive, circulatory, and genitourinary diseases, with an accuracy of c. 60-70%, it performs poorly for suicides, tuberculosis, and cancer. However, for tuberculosis and cancer, the algorithm is very good at assigning a code within the same Heiberg group as the validation data, even if the individual code for the unique cause of death is wrong. For tuberculosis, the accuracy is almost 100% at the group level, compared to 'just' 40% at the code level. Similarly, cancer diseases, for which the algorithm performs very poorly at the code level, are coded to the right Heiberg group almost 90% of the time. Finally, we find that the cause of death pattern based on the codes assigned by the string comparison method is fairly similar to the pattern based on the manual codes, which is impressive since the string comparison method only assigns a correct code to 46.3% of the unique cause of death strings. However, upon closer inspection there are several discrepancies between the two patterns which would result in two quite different results when analysing the cause of death patterns and their development over time. It seems most likely that these discrepancies have occurred because of the high percentage of wrongly assigned codes amongst the unique causes of death, and the fact that certain Heiberg groups are more affected by this than others.

From examining the automatic coding of the unique cause of death strings, we can point out two major flaws of the string similarity method. Firstly, a number of the relatively straightforward causes of death at the top of the frequency list were coded inaccurately according to the validation data simply due to significant differences in the wording between the two strings. Examples include "Kramper" ["cramps"], "ukendt" ["unknown"] and "pludselig død" ["sudden death"]. These issues could be addressed by adding synonyms to the coding scheme, so that it has more than one label for each category, and thus would be more flexible to the different but semantically identical terms. Secondly, in other cases, the automated method fails because the unique cause of death string in question fits into several categories in the DK1875 system. It is difficult to increase the accuracy of the string comparison method in these cases, since they require contextual knowledge. Even a manual coder will often need additional help such as clinical dictionaries and medical literature to know more about the range of distinct expressions of the same disease.

When we consider these issues, the version of the method applied here is not accurate enough to use for actual data analysis of the cause of death patterns. While it is encouraging that the string similarity comparison is close to presenting a somewhat accurate representation of the overall distribution cause of death groups, we could paint the same picture using aggregate cause-of-death statistics, which defeats the purpose of working on individual-level data. We do find that while some of the causes that were inaccurately coded were in fact coded to another category within the same Heiberg group, meaning that it was coded to an adjacent illness, others were coded to completely different categories. This would present a major issue if the data was used to study the development of specific causes. We would argue that scholars are increasingly interested in doing just that: analyses of specific illnesses, in connection to or rather than the overall panorama of disease.

The promise of automated coding is that it can help cover the diversity of causes, by coding the long tail of cause of death strings that only appear once or twice in the dataset, which a manual coder will rarely reach. Based on the method's performance on the validation data, we would not trust it to work well on a large dataset with no validation checks. However, it should be taken into account that the algorithm used was developed quickly during a 48-hour hackathon. Our results represent a developer effort that is realistic for individual research projects, and our code could be adjusted for other projects. The results of the paper provide a baseline, and it is very well possible to increase its accuracy. A way forward could be to work with the cut-off points of the Jaro-Winkler scores, setting a threshold so that only those causes where the string similarity match is relatively certain are coded and leaving the remainder to a manual coder. While doing so would ensure a higher degree of certainty in the code assigned, it would reduce the number of causes of death coded. In addition, the amount of uncoded causes of death would most likely not be equally distributed amongst all cause of death groups, meaning that even though the certainty of each code is higher, the sample in total may be more biased. As it is now, the automated coding would perhaps be better used as a method to catch most of the initial cases of a certain disease with a very set phrasing, such as cancer. Afterwards, the researcher could look for more cases, such as tumours. In this way, it would be a helpful tool for locating cases of the specific disease for a study of this disease. However, this only works for certain types of diseases, where the word is consistently used in both the coding and the unique cause of death string.

Acknowledgments

Many thanks to the Copenhagen City Archives and their volunteers for the creation of the Copenhagen Burial Register Dataset, without which none of this would have been possible.

References

[1] A. R. Omran, The epidemiologic transition: a theory of the epidemiology of population change., The Milbank Quarterly 83 (2005) 731–757. doi:10.1111/j.1468-0009.2005.00398.x.

[2] T. McKeown, The modern rise of population, Academic Press, Nueva York, 1976.

[3] A. Løkke, Døden i barndommen: spædbørnsdødelighed og moderniseringsprocesser i Danmark 1800 til 1920, Gyldendal, København, 1998.

[4] A. Reid, E. Garrett, C. Dibben, L. Williamson, 'A confession of ignorance': deaths from old age and deciphering cause-of-death statistics in Scotland, 1855–1949, The History of the Family 20 (2015) 320–344. URL: https://doi.org/10.1080/1081602X.2014.1001768. doi:10.1080/1081602X.2014.1001768, number: 3.

[5] J. P. Mackenbach, The Epidemiologic Transition Theory, Journal of Epidemiology and Community Health (1979-) 48 (1994) 329–331. URL: http://www.jstor.org/stable/25567930, publisher: BMJ.

[6] G. C. Alter, A. G. Carmichael, Classifying the Dead: Toward a History of the Registration of Causes of Death, Journal of the History of Medicine and Allied Sciences 54 (1999) 114–132. URL: http://www.jstor.org/stable/24624555.

[7] Det Statistiske Bureau, Statistisk Tabelværk, Fjerde række, Litra A Nr. 2, Vielser, Fødsler og Dødsfald i Aarene 1875-1879 samt Dødsaarsagerne i aarene 1876-1879., Bianco Lunos Bogtrykkeri, Kjøbenhavn, 1882.

[8] G. B. Risse, Cause of death as a historical problem, Continuity and Change 12 (1997) 175–188. URL: http://www.cambridge.org/core/journals/continuity-and-change/article/cause-of-death-as-a-historical-problem/74166C12EDE0C2B9C9AD025DD85948E4. doi:10.1017/S0268416097002890, number: 2.

[9] G. C. Alter, A. G. Carmichael, Reflections on the classification of causes of death, Continuity and change 12 (1997) 169–173. doi:10.1017/S0268416097002889, place: Cambridge Publisher: Cambridge University Press.

[10] A. Janssens, I. Devos, The Limits and Possibilities of Cause of Death Categorisation for Understanding Late Nineteenth Century Mortality, Social History of Medicine 35 (2022) 1053–1063. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9949560/. doi:10.1093/shm/hkac040.

[11] J. Arrizabalaga, Medical Causes of Death in Preindustrial Europe: Some Historiographical

Considerations, Journal of the History of Medicine and Allied Sciences 54 (1999) 241–260. URL: http://www.jstor.org/stable/24624562.

[12] S. J. Kunitz, Premises, Premises: Comments on the Comparability of Classifications, Journal of the History of Medicine and Allied Sciences 54 (1999) 226–240. URL: http://www.jstor.org/stable/24624561.

[13] G. C. Alter, A. G. Carmichael, Studying causes of death in the past : problems and models, Historical methods 29 (1996) 44–48.

[14] A. Janssens, Constructing SHiP and an International Historical Coding System for Causes of Death, Historical Life Course Studies 10 (2021) 64–70. URL: https://hlcs.nl/article/view/9569. doi:10.51964/hlcs9569.

[15] Københavns begravelsesprotokoller 1861–1911 (n=306.541) [The Copenhagen burial register 1861–1911], October 2020. URL: https://www.rigsarkivet.dk/udforsk/link-lives-data/, type: dataset.

[16] P. Harteloh, The implementation of an automated coding system for cause-of-death statistics, Informatics for health & social care 45 (2020) 1–14. doi:10.1080/17538157.2018.1496092, place: England Publisher: Taylor & Francis.

[17] J. Carson, G. Kirby, A. Dearle, L. Williamson, A. Reid, C. Dibben, "Exploiting historical registers: Automatic methods for coding 19th and 20th-century cause of death descriptions to standard classifications"., in: New Techniques and Technologies for Statistics, Eurostat, 2013, pp. 598–607.

[18] B. Koopman, G. Zuccon, A. Nguyen, A. Bergheim, N. Grayson, Automatic ICD-10 classification of cancers from free-text death certificates, International Journal of Medical Informatics 84 (2015) 956–965. URL: https://www.sciencedirect.com/science/article/pii/S1386505615300289. doi:10.1016/j.ijmedinf.2015.08.004.

[19] K. Xu, M. Lam, J. Pang, X. Gao, C. Band, P. Mathur, F. Papay, A. K. Khanna, J. B. Cywinski, K. Maheshwari, P. Xie, E. P. Xing, Multimodal Machine Learning for Automated ICD Coding, in: Proceedings of the 4th Machine Learning for Healthcare Conference, PMLR, 2019, pp. 197–215. URL: https://proceedings.mlr.press/v106/xu19a.html, iSSN: 2640-3498.

[20] K. Stadsarkiv, Begravelser i hele København 1861-1942, 2022-05-02. URL: https://kbharkiv.dk/brug-samlingerne/kilder-paa-nettet/begravelser-i-koebenhavn/begravelser-1861-og-frem/.

[21] L. Ludvigsen, B. Revuelta-Eugercios, A. Løkke, Cause-Specific Infant Mortality in Copenhagen 1861–1911 Explored Using Individual-Level Data, Historical Life Course Studies 13 (2023) 9–43. URL: https://hlcs.nl/article/view/12032. doi:10.51964/hlcs12032.

[22] B. Revuelta-Eugercios, H. Castenbrandt, A. Løkke, Older rationales and other challenges in handling causes of death in historical individual-level databases: the case of Copenhagen, 1880–1881, Social history of medicine : the journal of the Society for the Social History of Medicine (2021). doi:10.1093/shm/hkab037.

[23] K. Stadsarkiv, Indtastningsvejledning - Begravelser 1861-1940, 2023-03-17. URL: https://kbharkiv.dk/deltag/indtast-begravelser-1861-1940/indtastningsvejledning-begravelser-1861-1940/.

[24] B. Johansson, Den danske sygdoms- og dødsaarsagsstatistik : Med et afsnit om pneumonistatistik., Ejnar Munksgaard, Kbh, 1946.

[25] Anvisning for Læger med Hensyn til Udstedelsen af Dødsattester., Det Kongelige Sundheds-

Collegium, 1875. URL: http://www5.kb.dk/e-mat/dod/130021307433.pdf.
[26] J.-B. Ratte, pyjarowinkler 1.8, March 23, 2016. URL: https://pypi.org/project/pyjarowinkler/.
[27] P. Heiberg, Dødeligheden og Dødsaarsagerne i Danmark i de 2 Tiaar 1890-1899 og 1900-1909 i Aldersklasserne 15-74 Aar., Særtryk af Bibliotek for Læger, 1918.

A. Appendix A: The DK1875 system

Nr.	Latin	Danish
I	Morbi epidemici	Farsoter
1	Variolæ	Kopper
2	Morbilli	Mæslinger
3	Scarlatina	Skarlagensfeber
4	Diphtheritis	Ondartet Halssyge
5	Croup	Strubehoste
6	Tussis convulsiva	Kighoste
7	Febris tyhoidea	Typhoid Feber
8	Typhus exanthematicus	Exanthematisk Typhus
9	Dysenteria	Blodgang
10	Cholera asiatica	Asiatisk Cholera
11	Cholerine & Catarrhus intest. acutus	Indenlandsk Cholera og akut Diarrhoe
12	Erysipelas faciei & ambulans	Ansigts- og anden Vandrerosen
13	Febris puerperalis	Barselfeber
14	Pyæmia & Septichæmia	Ondartet Saarfeber
15	Febris intermittens	Koldfeber
16	Influenza	Grippe
17	Febris rheumatica	Akut Ledderheumatisme
18	Alii morbi epidemici	Andre Farsoter
II	Sangvinis infectiones	Blodforgiftninger
19	Malleus humidus	Snive
20	Pustula maglina	Miltbrand
21	Alia venena animalia	Andre dyriske Gifte
22	Syphilis acquisita	Erhvervet Syphilis
23	Syphilis congenita	Medfødt Syphilis
24	Alcoholismus chronicus	Brændeviinssygdom
25	Delirium tremens	Drankergalskab
26	Mors in ebrietate	Pludselig Død af Drik
III	Morbi Constitutionales	Konstitutionelle sygdomme
27	Scrophulosis	Kirtelsyge
28	Hydrocephalus acutus	Akut Hjernevandsot
29	Tuberculosis acuta	Akut Miliærtuberkulose
30	Phthisis pulmonum	Lungesvindsot

31	Tuberculosis in aliis corporis partibus	Tuberkulose i andre Organer
32	Cancer ventriculi	Mavekræft
33	Cancer uteri	Livmoderkræft
34	Cancer mammæ	Brystkræft
35	Cancer in aliis corporis partibus	Kræft i andre Organer
36	Rhachitis	Engelsk Syge
37	Diabetes mellitus	Sukkersyge
38	Scorbutus	Skjørbug
39	Anæmia	Anæmi
IV	Violentæ mortis causæ	Voldsomme dødsårsager
40	Casus Mortiferi: Præcipitatio & Contusio	Ulykkelige Hændelser: Fald og Knusning
41	Casus Mortiferi: Submersio	Ulykkelige Hændelser: Drukning
42	Casus Mortiferi: Suffocatio	Ulykkelige Hændelser: Kvælning og Ihjelliggen
43	Casus Mortiferi: Vulnus sclopetarium	Ulykkelige Hændelser: Skudsaar
44	Casus Mortiferi: Vulnus incisum & punctum	Ulykkelige Hændelser: Snit- og Stiksaar
45	Casus Mortiferi: Ambustio	Ulykkelige Hændelser: Forbrænding og Skoldning
46	Casus Mortiferi: Congelatio	Ulykkelige Hændelser: Forfrysning
47	Casus Mortiferi: Veneficium	Ulykkelige Hændelser: Forgiftning
48	Casus Mortiferi: Alii casus mortiferi	Ulykkelige Hændelser: Andre ulykkelige hændelser
49	Suicidium: Submersio	Selvmord: Drukning
50	Suicidium: Strangulatio	Selvmord: Hængning
51	Suicidium: Vulnus sclopetarium	Selvmord: Skud
52	Suicidium: Vulnus incisum & punctum	Selvmord: Snit og stik
53	Suicidium: Venenum	Selvmord: Gift
54	Suicidium: Alii suicidii modi	Selvmord: Andre Selvmords Dødsaarssager
55	Homicidum	Mord og Drab
V	Vitia innata	Dannelsesfejl
56	Cyanosis	Blaasot
57	Debilitas congenita	Medfødt Svaghed
58	Spina bifida	Medfødt Rygmarvsvandsot
59	Atelectasis	Mangelfuld Udvidning af Lungerne
60	Alia vitia innata	Andre Dannelsesfejl
VI	Morbi singulorum organorum	Lokale organsygdomme
61	Encephalitis et Meningitis cerebralis	Hjernebetændelse
62	Apoplexia cerebri	Apoplexi
63	Morbi cerebri chronici	Chroniske Hjernesygdomme
64	Morbus mentalis	Sindssygdom
65	Tetanus	Stivkrampe
66	Trismus	Mundklemme
67	Epilepsia	Ligfald
68	Ecclampsia	Konvulsioner

69	Myelitis & Meningitis spinalis	Rygmarvsbetændelse
70	Ataxia s. Tubes dorsalis	Rygmarvstæring
71	Alii medullæ spinalis morbi chronici	Andre chroniske Rygmarvssygdomme
72	Laryngitis	Strubebetændelse
73	Morbi laryngis chronici	Chroniske Strubesygdomme
74	Pneumonia	Lungebetændelse
75	Pleuritis, Empyema	Lungehindebetændelse
76	Bronchitis acuta simplex	Brystkatarrh, akut bronchitis
77	Bronchitis capill. & Pneumonia catarrh	Kapillær Bronchitis og katarrhalsk Lungebetændelse
78	Bronchit. chron. & Bronchiectasis	Chronisk Bronchitis
79	Emphysema pulmonum & Emphysem, Astma	
80	Alii pulmonum morbi chronici	Andre chroniske Lungesygdomme
81	Peri- & Endocarditis	Betændelse af Hjertet og dets hinder
82	Morbus Cordis	Organisk Hjertesygdom
83	Aneurysma Aortæ	Udvidning af Aorta
84	Phlebitis	Blodaarebetændelse
85	Ulcus perforans ventriculi	Perforerende Mavesaar
86	Peritonitis	Bughindebetændelse
87	Enteritis, Colitis, Typhlitis	Tarmbetændelse
88	Ileus	Tarmslyngning
89	Hernia incarcerata	Indeklemt Brok
90	Cirrhosis hepatis	Lever-Cirrhose
91	Echinococcus hepatis	Lever-Echinokok
92	Cholelithiasis	Galdesten
93	Nephritis albuminosa	Brights Sygdom
94	Lithiasis renalis & vesicalis	Nyre- og Blæresten
95	Cystitis	Urinblærebetændelse
96	Strictura urethræ	Forsnevring af Urinrøret
97	Hypertrophia prostatæ	Chronisk Prostatasygdom
98	Tumor ovarii, Hydrops ovarii	Æggestok-Svulst
99	Alii morbi abdominales chronici	Andre chroniske Underlivssygdomme
100	Hydrops ex ignota causa ortus	Vandsot af ubekjendt Aarsag
101	Alii morbi organorum interiorum	Andre Sygdomme i indvendige Organer
VII	Morbi externarum partium	Sygdomme i de ydre dele
102	Phlegmone, Abscessus	Bindevævsbetændelse
103	Caries & Necrosis ossium	Benedder
104	Arthrocace	Leddebetændelse
105	Fractura coli femoris	Brud af Laarbenets Hals
106	Gangræna	Koldbrand
107	Carbunculus & Furunculus malignus	Brandbyld
108	Alii externarum partium morbi	Andre Sygdomme i de ydre Dele
VIII	Aliæ causæ mortis freqventes	Andre hyppige Dødsårsager
109	Marasmus senilis	Alderdomssvaghed

110	Atrophia infantilis	Tæring hos Smaabørn
111	Mors in partu & puerperio (Fb. puerp. excl.)	Død under Fødslen og i Barselsseng (Barselfeber ikke medregn.)
112	Mors repentina sine nota causa	Pludselig Død uden bekjendt Aarsag
113	Causa mortis vel male vel omnino non indicata	Uangiven eller slet specificeret Dødsaarsag
114	Mors medico non vocato obveniens	Død uden Lægebehandling
IX	Exanimus natus	Dødfødte
115	Exanimis natus	Dødfødt

B. Appendix B: The Heiberg groups

Heiberg group nr.	Heiberg group name	DK-1875 categories
I	Infectious diseases	1-23
II	Croupous Pneumonia	74
III	Tuberculosis	27-31
IV	Cancer	32-35
V	Alcoholism	24-26
VI	Suicide	49-54
VII	Accidents	40-48
VIII	Diseases in the nervous system	61-71
IX	Respiratory diseases	72, 73, 75-80
X	Cardiovascular diseases	81-84
XI	Digestive diseases	85-92
XII	Genitourinary diseases	93-97
XIII	Others	36-39, 55-60, 98-115

Wikidata for authority control: sharing museum knowledge with the world

Alicia Fagerving[1]

[1] Wikimedia Sverige, c/o Norrsken House, Birger Jarlsgatan 57C, 113 56 Stockholm, Sweden

Abstract

The development of Wikidata as the web's central authority hub has created new opportunities for cultural heritage institutions wishing to make their data more visible, accessible and relevant. This paper introduces the project *Usable Authorities for Data-driven Cultural Heritage Research*, in which the Nationalmuseum and the National Historical Museums, in partnership with Wikimedia Sverige, explore this new area.

The aim of our project is to develop and evaluate methods of linking museum authority data to Wikidata, with the ultimate goal of making it easier for researchers and other users to find, understand and analyze relevant information distributed across different museum collections. The project includes visualizations of relations between historical persons and places to show the potential of otherwise abstract large amounts of data. Another aim is to increase the knowledge about Wikidata and its possibilities among cultural heritage institutions in Sweden, encouraging them to use the data and resources accumulated on the Wikimedia platforms, as well as to actively contribute to them themselves.

The project has proven to be an innovative case study on what is required from a cultural heritage institution to make their data more open and shareable. The focus is on cleaning up the authority data in the museums' databases, including correcting errors and deleting duplicates. The data is there aligned with Wikidata using OpenRefine, so that relevant Wikidata items can be updated with the museums' identifiers. Later on, by developing attractive visualizations, the data is brought to life, highlighting how much information Wikidata volunteers have contributed with and how it complements the museums' own knowledge bases. The visualizations showcase how Wikidata, by being heavily structured and machine-readable, enables the development of new applications that bring the data closer to the users.

Keywords

Wikidata, Wikimedia, authority control, Linked Open Data, GLAM, citizen humanities

1. Introduction

1.1. Linked Open Data and cultural heritage

While numerous cultural heritage institutions have been working on digitizing their collections and processes, in many cases this work has been done independently from one another. Due to this, there are differences not only in their maturity levels and the role of digital thinking in their strategies, but also in the specific tools, platforms and designs they have implemented, which creates thresholds for those who wish to access and re-use their data. This includes actors such as developers who build applications using data from multiple sources, data journalists,

DHNB2023 | Sustainability: Environment - Community - Data. The 7th Digital Humanities in the Nordic and Baltic Countries Conference. Oslo – Stavanger – Bergen, Norway. March 8–10, 2023.

alicia.fagerving@wikimedia.se (A. Fagerving)

DHNB Publications, DHNB2023 Conference Proceedings, https://journals.uio.no/dhnbpub/issue/view/875

researchers and the general public. A move from disparate data silos to shared platforms that enable and encourage participation of multiple actors makes the data more visible, accessible and re-usable. While there has been a growing interest in Linked Open Data, it has not yet become a common standard among cultural heritage institutions.[1]

The development of sustainable, interoperable and open digital infrastructures is one of the main focuses of the Digital Decade Policy Programme 2030, in which the EU Parliament, the EU member states and the European Commission together identify the key areas for the development and implementation of a shared European digital vision.[2] This is the context in which the work described in this paper is being done. When museum data becomes more available and easier to share, more people can access and learn from it in the spirit of democracy. The Wikimedia platforms, of which Wikipedia is the most known example, welcome everyone to participate and contribute on equal terms, regardless of their geographical location, their choice of language or formal qualifications.

A common hub for Linked Open Data, like Wikidata, plays a pivotal role in facilitating interoperability between different institutions' collections, which offers numerous opportunities for both the institutions themselves and those who consume their data – be it the general public, developers, the civil society or researchers. While many museums have been working on making their knowledge more digital and available over the years, they have all established their own systems, structures and processes. Discovering relationships, similarities and differences between different institutions' datasets becomes easier with the help of Wikidata, which this paper will show with a practical example.

1.2. Wikidata – an open data platform for everyone

Wikidata is a collaborative, multilingual Linked Open Data (LOD) platform, launched by the Wikimedia Foundation in 2012. It is a sister project of Wikipedia, and just like Wikipedia it can be accessed and edited by anyone.

While it was launched with the purpose of serving as a central data management platform for Wikipedia,[3] with its growing popularity and increasing use in various fields, Wikidata is becoming an important resource for researchers, developers, and other stakeholders who seek to leverage structured data for a variety of applications. The potential of Wikidata for connecting collections and performing authority control has been noted by Europeana, who in 2015 released a set of recommendations centered around an increased inclusion of the Wikimedia platforms, in particular Wikidata, in the work of GLAM (Gallery, Library, Archive and Museum) institutions.[4] For Digital Humanities applications, Wikidata is a valuable knowledge base as it contains data on a wide range of topics while making it easy to crowdsource information and providing a flexible platform for enrichment and dissemination of data.[5]

Wikidata's approach to data management ensures that the data is not only easily accessible to human readers, but also machine-readable, making it available to scientists, software developers, data journalists and the like. The fact that data on Wikidata is available under the Creative Commons CC0 License contributes to the easiness of re-use, as it does not require attribution. However, it also means that only datasets free from copyright restrictions can be mass ingested into Wikidata, with the consequence that organizations that release truly open data have a head start when contributing to the platform, which as of March 2023 contains data on over

102 million entities.[1]

1.3. Wikidata as an authority hub

Each entity in Wikidata can contain links to external databases ("external identifiers"), allowing for easy cross-referencing of information. Since its inception, Wikidata has organically grown into a Linked Open Data hub, with thousands of properties for identifiers in databases such as VIAF, GeoNames and the Library of Congress.[6] Authority databases represented in Wikidata range from global and general (over 3 million items are linked to VIAF) to national and domain-specific, such as the Swedish Literature Bank edition ID (P5123),[2] which links items of book editions with their corresponding entries in the digital library Litteraturbanken.[3]

In the context of the work presented here, namely museums in Sweden, there are several museums whose identifiers can be used in Wikidata, including the Nationalmuseum and the National Historical Museums of Sweden. In addition, items can be linked to Swedish Open Cultural Heritage (K-samsök), which aggregates metadata from over 50 cultural heritage institutions.[4]

1.4. Starting the journey towards linked authority data

Wikimedia Sverige (WMSE) is a Swedish non-profit organization that promotes the use and development of the Wikimedia platforms, including Wikipedia, Wikidata and Wikimedia Commons. WMSE provides education and training in contributing to these platforms, and supports GLAM institutions in digitizing and sharing their collections through them. Over the years, WMSE has supported a variety of institutions such as the Nordic Museum, the National Library of Sweden and the Swedish Performing Arts Agency in their efforts to share their resources on the Wikimedia platforms. Usually, this takes the form of a collaborative project where the institution provides a set of digitized images or data that WMSE then makes available on Wikimedia Commons or Wikidata. While the selection of the material and decisions on the best way of structuring it on the Wikimedia platforms are made together, WMSE often takes on a leading role as an expert in the free knowledge ecosystem.

In this project, Wikimedia Sverige is working together with the **Nationalmuseum** and the **National Historical Museums**. They have both worked with the Wikimedia platforms and WMSE previously: WMSE has supported them in sharing thousands of digitized artworks and photographs on Wikimedia Commons, the central repository of free media files in the Wikimedia ecosystem. Because of this, there was already an awareness of the Wikimedia platforms and their role in the free knowledge ecosystem, as well as a curiosity about more advanced applications. For at least half a decade, the two museums have contributed valuable material to the Wikimedia platforms, as well as received feedback about the impact of their contributions on readers and the editor community alike. They have also included the Wikimedia platforms in their own digital presence – the Nationalmuseum entries on their artworks link to the high resolution

[1] "Wikidata Main Page". Accessed March 15, 2013. https://www.wikidata.org/wiki/Wikidata:Main_Page.

[2] "Property talk:P214". Wikidata. Accessed March 17, 2023. https://www.wikidata.org/wiki/Property:P214.

[3] "Litteraturbanken". Svenska Akademien. Accessed March 18, 2023. https://litteraturbanken.se/.

[4] "Template:Sweden properties". Wikidata. Published November 26, 2022. Accessed March 18, 2023. https://www.wikidata.org/w/index.php?title=Template:Sweden_properties&oldid=1778976922.

versions on Wikimedia Commons.[5]

2. Project outline

The project *Usable Authorities for Data-driven Cultural Heritage Research* is a collaboration between the Nationalmuseum, the National Historical Museums and Wikimedia Sverige, running from 2021 to 2023. It is financed by a Swedish National Heritage Board grant via their Research and Development program. Wikimedia Sverige has previously worked with the Swedish National Heritage Board, who have a long track record of both publishing and supporting open data in the cultural heritage sector.

The goal of the project is to investigate whether vocabularies as linked data stimulate new data-driven research on museum collections and whether Wikidata is a suitable platform to use to link different authorities and vocabularies together. This is achieved by developing a method to make Swedish personal authority files compatible with each other, i.e. by increasing the semantic interoperability of museum collections. This method should be documented, tested and possible to adjust to the needs of other museums. The project applies the FAIR principles for data management: the data must be Findable, Accessible, Interoperable and Reusable.[7]

- Publishing all of the National Historical Museums' personal authority files online, with persistent identifiers for every personal data post.
- Linking all personal data entries of people born between 1500 and 1800 in the Nationalmuseum and the National Historical Museums collections to Wikidata.
- Publishing at least 5 visualizations of the data that demonstrate the power of Linked Open Data.
- Producing a qualitative appraisal of whether the interlinking of authority data in the project is beneficial for research in the participating museums.
- Spreading awareness about LOD and Wikidata in the museum sector and in the research community.

3. Implementation

Apart from the concrete work of sharing the data on the Wikimedia platforms, a long-term goal of the project is to equip the team members with the tools and skills needed to continue working with Wikidata in a strategic, independent way. For this reason, the first stage of the project was extensive training provided to the two museums' team members by WMSE. The team members were encouraged to edit Wikidata manually before introducing tools and strategies for mass editing, and an emphasis was placed on understanding the community-developed conventions and best practices of the platforms.

WMSE provided practical assistance during the initial stages of the data uploads, so that the team members could see an example of how a large-scale upload might be done. However,

[5] See for example https://collection.nationalmuseum.se/eMP/eMuseumPlus?service=direct/1/ResultDetailView/result.tab.link&sp=10&sp=Scollection&sp=SfieldValue&sp=0&sp=2&sp=3&sp=SdetailView&sp=11&sp=Sdetail&sp=0&sp=F&sp=SdetailBlockKey&sp=5.

emphasis was again placed on development of independent skills. Rather than having the team members submit metadata files for WMSE to be uploaded, they were encouraged to work on their own in small batches and communicate with each other – across the two museums – in order to help each other identify and solve problems. The WMSE team members were present in the regular check-ins in order to keep track of the progress and discuss questions that required more experience of the Wikimedia platforms.

At the end of Year 1 and Year 2, seminars were organized for the Swedish cultural heritage community, where they were both updated about the progress of the project, including the challenges the team was facing, and shown practical demonstrations of Wikidata and OpenRefine. The seminars were received positively, with requests for more exhaustive events of this type, which is why Year 3 will see more emphasis on outreach and documentation. As one of the goals of the project is to show other museums what can be achieved with the help of Wikidata, we will intensify our efforts to spread our outcomes and learnings.

In April 2023, we hosted an all-day workshop for museum professionals where they tried their hand at both working with domain-specific data on Wikidata and analyzing and processing datasets in OpenRefine. It has been our experience that GLAM professionals in Sweden have basic knowledge about Wikidata and are curious about its potential, but need assistance designing and implementing practical projects – a need that we directly addressed in the workshop.

Furthermore, we are planning to publish a handbook for museum professionals in which we will gather the experiences and learnings from the project. It will include the perspectives of the Nationalmuseum and the National Historical Museums as well as of Wikimedia Sverige in order to paint a comprehensive picture of a longer collaborative project between a GLAM and a Wikimedia chapter, its prerequisites, outcomes and pitfalls.

Finally, a set of data visualizations was prototyped early in the project, and continued to develop in parallel with the progress of the data uploads.

4. Tools and skills

In order to be able to use and contribute to Wikidata in an efficient way, the working group had to learn a number of skills and tools. These are in no way unique to this project, but will continue to benefit the participants shall they wish to implement the Wikimedia platforms in other areas of their work.

As Wikidata is a community-driven project, its structures and data models have grown organically over the years, with the input of thousands of volunteers from all over the world. Not only the content, but also the ontology itself is created collaboratively and steadily evolving.[8] It should be noted that everyone contributes to Wikidata on equal footing, regardless of whether they are doing it as part of their professional activities or as volunteers; decisions are made collectively based on consensus.[6] Because of this, the early training was to a large extent based on examining the cultural heritage data in Wikidata (persons, objects, institutions) and the ways the community around it organizes itself; for example the *Wikiproject Sum of All Paintings*,[7]

[6] "Consensus". Meta-Wiki. Published August 15, 2021. Accessed March 15, 2023. https://meta.wikimedia.org/w/index.php?title=Consensus&oldid=21883628.

[7] "WikiProject Sum of All Paintings". Wikidata. Published March 6, 2023. Accessed March 15, 2023. https:

which aims to create comprehensive items for all notable paintings in the world and is an area where the Nationalmuseum, especially, can contribute a lot of content. Our work was made considerably easier by the fact that the project team had worked with Wikipedia and Wikimedia Commons previously, thanks to which they were already familiar with the underlying principles of the Wikimedia platform, such as collaborative editing and consensus.

As the project encompasses editing and creating large numbers of Wikidata items, it was crucial to identify a suitable tool for mass editing. Several options are available, which differ greatly in terms of complexity and the technical skills required. Programmers can write their own tools using one of several available libraries, such as Pywikibot[8] for Python or Wikidata Toolkit[9] for Java. QuickStatements[10] is a volunteer-developed web-based tool that makes it possible to batch edit data from, for example, a spreadsheet. While QuickStatements makes it easy to quickly add data, such as external identifiers, to a set of items, it requires the user to have identified the items to be edited. In our project, our first need was to be able to automatically match the strings exported from the museums' databases, such as the names of people and places, to corresponding Wikidata items.

This need is met by **OpenRefine** (OR).[11] Formerly known as Google Refine, OR is a powerful software tool for data cleaning and transformation. It provides users with a spreadsheet-like interface to analyze and manipulate datasets, enabling them to perform a wide range of data cleaning tasks such as data parsing, filtering, and editing. Thanks to the built-in Wikidata integration, users can edit and create items directly from the application, aligning their data structure to Wikidata in a visual schema builder that looks not dissimilar to the Wikidata interface. Custom operations can be done using either GREL (General Refine Expression Language),[12] which resembles JavaScript, Jython or Closure. OpenRefine is especially useful when working with large datasets with a high degree of variability in their formats or values. The tool supports various data sources, including CSV, TSV, JSON, XML, and Google Sheets, making it highly flexible for data processing tasks. Most importantly, it has powerful reconciliation functionalities, allowing its users to match their datasets with external sources, including but not limited to Wikidata.[13] The software is cross-platform, and the user can choose from running it on their own computer or using the fully functional remote instance hosted on the Wikimedia Cloud Services.[14]

OpenRefine has become a popular tool in data science and analysis communities. The tool is open source and is being developed actively; WMSE has close contact with the developer team,

//www.wikidata.org/w/index.php?title=Wikidata:WikiProject_sum_of_all_paintings&oldid=1846242883.

[8] "Wikimedia/Pywikibot". GitHub. Published January 9, 2023. Accessed March 15, 2023. https://github.com/wikimedia/pywikibot.

[9] "Wikidata/Pywikibot". GitHub. Published March 9, 2023. Accessed March 15, 2023. https://github.com/Wikidata/Wikidata-Toolkit.

[10] "QuickStatements". Accessed March 15, 2023. https://quickstatements.toolforge.org/.

[11] https://openrefine.org/.

[12] "General Refine Expression Language". OpenRefine. Published November 24, 2022. Accessed March 16, 2023. https://openrefine.org/docs/manual/grel.

[13] "Reconciling". OpenRefine. Published January 6, 2022. Accessed March 16, 2023. https://openrefine.org/docs/manual/reconciling.

[14] "PAWS". Wikitech. Published February 23, 2023. Accessed March 15, 2023. https://wikitech.wikimedia.org/w/index.php?title=PAWS&oldid=2055950.

making it easy for us to inform the users about new features and pass on bug reports. This, in connection with the fact that OR has extensive documentation, has made it an obvious choice for this project. OR has been successfully used by GLAMs for both Wikidata-focused projects[9] and internal data cleaning tasks,[10][11] creating a springboard of ideas and experiences to learn from. Documentation and tutorials aimed specifically at the GLAM sector are available.[12][12]

Data can be retrieved from Wikidata using SPARQL, an RDF query language. In addition to a SPARQL endpoint, Wikidata provides a GUI query service (WDQS),[15] which is being used extensively by the team members to both gain insights into relevant data (e.g. the gender distribution of the artists with works in the collections of the Nationalmuseum) and to find errors and areas for improvement (e.g. how many of the artists lack a date of birth). On the one hand, learning SPARQL might seem intimidating to users without a programming background; on the other hand, even simple queries are a powerful tool to get an overview of the data in Wikidata, which is why a heavy emphasis was put on SPARQL in the early training. WDQS contains an extensive library of sample queries which can be explored and modified, further reducing the learning curve. An additional benefit of learning SPARQL is that it gives its users access to the basic visualization functionalities that are built into the WDQS, where the query results can be displayed as a chart, or overlaid on a map.

The Wikimedia platforms in general, and Wikidata in particular, can be difficult for newcomers to learn, especially if they come with the intention to start editing right away. The documentation that has grown organically over the years is spread across many pages, and conventions that are obvious to seasoned editors. One example that was brought up in the training was that the value of a *located in the administrative territorial entity* statement for an item representing a geographical location in Sweden should always be a municipality item, not a city or village item. This makes it easy to make mistakes when editing in good faith, with the edits often being reverted or deleted by more experienced editors without explanation, leading to frustration and distrust in the platform. This is why assistance from WMSE was important both in the initial stage of the project and throughout, as conventions and common pitfalls could be explained as needed.

5. Challenges

5.1. Tools and platforms

While it is possible to access a variety of operations, such as column splitting and merging, in OpenRefine from its built-in menus, users gain a lot of flexibility and customizability by writing code snippets to process the data. It has been our experience that GLAM professionals who lack a programming background feel hesitant towards this task. However, in our project, we have found that the team members were able to both learn the elements of GREL needed for their particular assignment and to navigate its documentation. In fact, they have reported that they have started to use OR not only for Wikidata reconciliation and data upload, but also for the internal work of data cleaning and analysis. The software makes it easy for them to get an overview of large datasets and to spot and correct errors.

[15] "Wikidata Query Service". Wikidata. Accessed March 14, 2023. https://query.wikidata.org/.

One drawback of using OpenRefine are its limited collaboration capabilities. Only one person can work on a dataset at a time (unlike in web-based spreadsheet software like Google Sheets), which presented a problem when several team members from the same museum wanted to share their workload or request a colleague's assistance. A cumbersome half-measure to solve this problem is to export one's OR project, send it to someone else and then re-import the changed version. The need for collaborative editing has been noted by OR developers, but it is not being actively worked on at the moment.[16]

5.2. Reconciliation

The automatic reconciliation functionality in OpenRefine cannot always be relied upon. It works best when there is an exact match between the string being reconciled and exactly one Wikidata item, as is the case with persons with unique names. If there are multiple matches, it is also possible to further refine the reconciliation process so that it takes into account additional data, such as the country of citizenship, profession or date of birth. However, this requires that both the source data and the Wikidata item are detailed enough. We encountered particular problems with persons born in the lower range of our timeframe (close to 1500), as their names can come in several variants and exact birth and death dates might not be available. There are also cases where correct Wikidata items cannot be found automatically due to a spelling mismatch – if a Ukrainian person only has a Wikidata label in Ukrainian, while the source dataset only uses a Swedish transliteration, the software will not recognize they should be linked together. Another problematic area are place names, especially those of villages and farms. There can be several sharing the same name, including historical ones that no longer exist.

When automatic reconciliation is not available, the team members must do the manual work of reviewing reconciliation candidates and possibly doing additional research in the collections to verify or append the available information. This can be very time-consuming, especially if it involves consulting colleagues who are experts in particular domains. However, this is necessary in order to identify a corresponding item in Wikidata, or to assure that one does not exist and a new one can be created. While accidentally creating a duplicate item is not catastrophic – items can be merged – editors should go to the greatest extent possible to avoid it.

6. Results

6.1. Data uploads

6.1.1. National Historical Museums

The National Historical Museums of Sweden ID (P9495)[17] is an external identifier property that links items on Wikidata to their corresponding entities in the museum database. This property

[16] "Development Roadmap". OpenRefine. Published November 24, 2022. Accessed March 17, 2023. https://openrefine.org/docs/technical-reference/development-roadmap.

[17] https://www.wikidata.org/wiki/Property:P9495

Table 1
The topical distribution of Wikidata items linking to the National Museum database. March 2023.

Type of item	Number of items	Example
person or organization	10,000	Gustavus Adolphus of Sweden (Q52938)
geographical location	10,000	Ängelholm Municipality (Q255206)
term	3,500	church bell (Q3500690)
taxon	1,200	Gasterosteiformes (Q212492)
object	820	Elizabeth Reliquary (Q26253636)
event	100	Battle of Öland (Q2036383)

did not exist before the start of the project and was created in April 2021 at the request of the project team.[18] It proved to be an opportunity for the team to learn about the property creation process. At first, the property was called "National Historical Museums of Sweden agent ID" and only used for items about persons and organizations. Later on, the museum decided to expand the scope of their data sharing to also include geographical locations, historical events and other types of items. This spurred a community discussion on whether additional properties should be created, or whether the scope of the P9495 property should be extended.[19] After consulting with the community, the latter was decided upon, as it is both simpler and more future-proof: there would be no need to create additional properties later down the line, should the need arise. As of March 2023, the National Historical Museums' identifier has been added to approximately 26,000 Wikidata items. It should be kept in mind that some of these edits have been made by the Wikidata community, as is to be expected and encouraged when working on an open and collaborative platform. The majority of them are items of persons and organizations, as well as geographical locations, but terms, taxa, objects and historical events are also represented. See Table 1 for the topical distribution of the items.

Simultaneously with adding the identifiers to Wikidata, links to Wikidata were also added to the corresponding entries in the museum's database. These connections are now publicly visible through in the entries on the website of the collections.

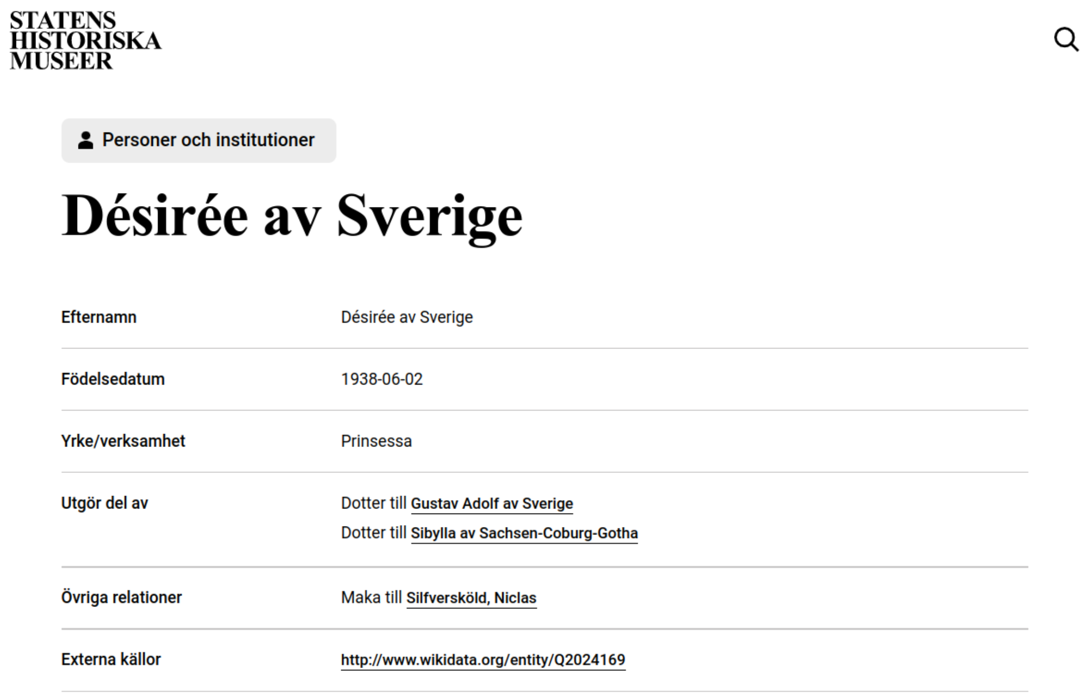
6.1.2. Nationalmuseum

The Nationalmuseum Sweden ID (P2538)[20] is an external identifier property that links items on Wikidata to their corresponding entities in the authority database of the Nationalmuseum. As of March 2023, the identifier has been added to over 10,000 Wikidata items. The property was created in 2016 and was already actively used before the project started. As much of the groundwork had already been done by the community, focus was put on refining existing entries, on importing the Wikidata identifiers to the museum database, as well as on creating links between the artists and their works on Wikidata.

[18] "Wikidata:Property proposal/National Historical Museums of Sweden agent ID". Wikidata. Published April 30, 2021. Accessed March 17, 2023. https://www.wikidata.org/w/index.php?title=Wikidata:Property_proposal/National_Historical_Museums_of_Sweden_agent_ID&oldid=1411325589

[19] "Property talk:P9495". Wikidata. Published April 21, 2022. Accessed March 18, 2023. https://www.wikidata.org/w/index.php?title=Property_talk:P9495&oldid=1622853192

[20] https://www.wikidata.org/wiki/Property:P2538



STATENS
HISTORISKA
MUSEER

Personer och institutioner

Désirée av Sverige

Efternamn	Désirée av Sverige
Födelsedatum	1938-06-02
Yrke/verksamhet	Prinsessa
Utgör del av	Dotter till Gustav Adolf av Sverige Dotter till Sibylla av Sachsen-Coburg-Gotha
Övriga relationer	Maka till Silfversköld, Niclas
Externa källor	http://www.wikidata.org/entity/Q2024169

Figure 1: The National Historical Museums link to Wikidata from their own collection website (https://samlingar.shm.se/person/96C81000-D580-4740-A6AB-E31057823DAC)

Wikidata has another external identifier property related to the Nationalmuseum's collections, Nationalmuseum Sweden artwork ID (P2539).[21] As of March 2023, there are at least 39,000 items of artworks that make use of it, some of which are a result of earlier collaborations between the museum and WMSE. As the museum database contains information about the people and locations depicted in some of the artworks, and Wikidata makes it possible to include this information in artwork items, much of Nationalmuseum's work in the project focused on sharing it in Wikidata. As a result, as of March 2023, the property Depicts (P180) is used in approximately 6,500 items of Nationalmuseum artworks[22], with 9,000 distinct objects, people and locations depicted.[23]

Additionally, data on approximately 800 exhibitions held by the Nationalmuseum from its inception until the 1970's has been shared on Wikidata.[24] Over 6,700 artwork[25] items have been linked to those using the property Exhibition history (P608).[26] The time range covered is a

[21] https://www.wikidata.org/wiki/Property:P2539
[22] https://w.wiki/6TmH
[23] https://w.wiki/6TmM
[24] https://w.wiki/6TmP
[25] https://w.wiki/6TmY
[26] https://www.wikidata.org/wiki/Property:P608

Table 2
The topical distribution of Wikidata items linking to the Nationalmuseum database. March 2023

Type of item	Number of items	Example
artwork	39,000	Midwinter's Sacrifice (Q761681)
person	10,000	Viveka Nygren (Q18274644)
exhibition	800	*Handteckningar av äldre danska målare* (Q109467267)

result of the limited availability of more current structured data; however, there is a possibility of obtaining and sharing more up-to-date exhibition data in the future.

6.2. Visualizations

Simply sharing data on Wikidata is not the only goal of this project. As Wikidata provides structured, machine-readable information, it is possible to re-use the data in a variety of ways. By visualizing the data, something that has a long tradition in scholarly education and communication and has been done by others working with cultural heritage data on Wikidata,[13] our goal was twofold. Firstly, we wanted to bring the newly shared data and the richness of Wikidata to the attention of both the general public and other GLAM professionals. While Wikidata itself is a niche platform, requiring a certain level of interest and expertise to appreciate, visualizations are much more accessible to Internet users. We also hoped to see links and connections between the people and the objects in the two museums' connections that were not immediately obvious.

In order to achieve this, a contractor with an experience of working with cultural heritage data on the Wikimedia platforms was brought in and created a set of visualizations in which data from the two museums is mixed with the community-generated data on Wikidata.[27] Below are three examples that demonstrate different aspects of our work.

Wikidataeffekten (The Wikidata Effect) compares the amount of data about an artist contained in Wikidata, in the National Historical Museums and in the Nationalmuseum.[28] It also fetches the other external identifiers from the artist's item in order to show which other institutions all over the world have an entry about the artist in their own authority databases. In this way, the sheer amount of data in Wikidata, as well its role as an authority hub, are highlighted.

Visbykartan (Map of Visby) combines Wikidata with OpenStreetMap, a collaborative mapping project, in order to show the people and artworks in the collections of the museums linked to places of interest in the medieval town of Visby.[29] This shows the potential of using multiple open data sources to build something that is more than merely a sum of its parts.

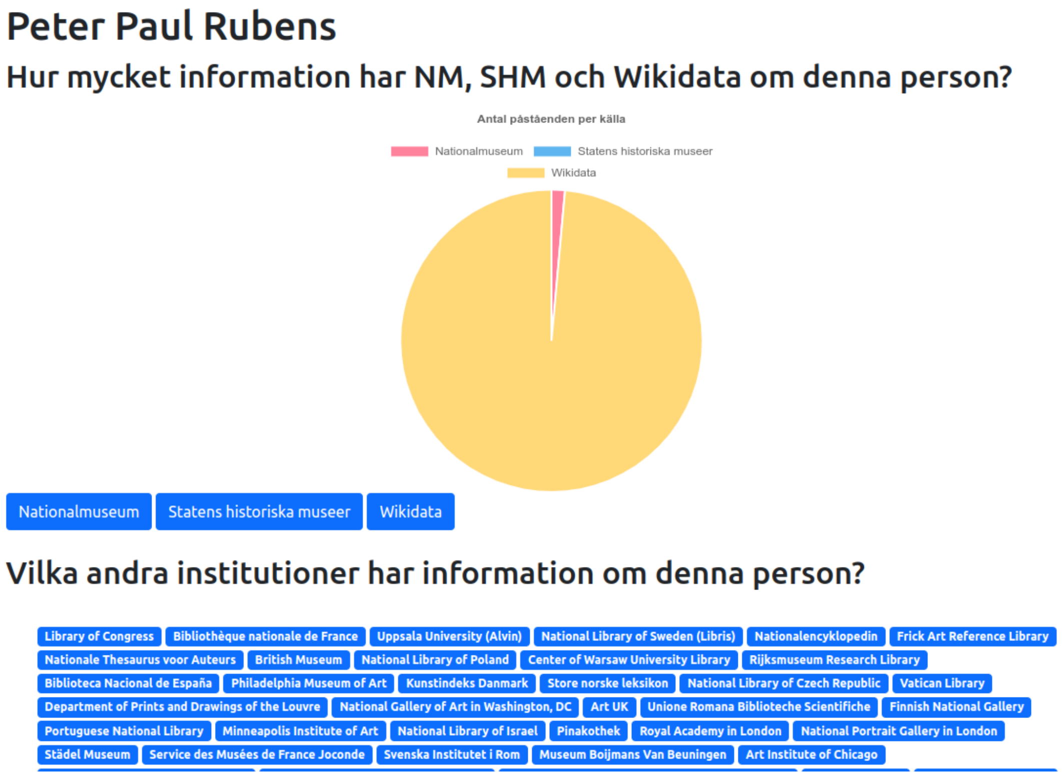
Konstnärsliv (Artists' Lives)[30] presents the user with an interactive map on which lines are drawn between the places of birth and death of artists represented in the Nationalmuseum. A slider makes it possible to see trends and patterns in how artists have moved away from their

[27] Larsson, Albin. "Visualiseringar för datadriven samlingsforskning". Accessed March 14, 2023. https://byabbe.se/datadriven-samlingsforskning/.

[28] See for example Peter Paul Rubens https://byabbe.se/datadriven-samlingsforskning/effekt/?qid=Q5599

[29] https://byabbe.se/datadriven-samlingsforskning/visby/

[30] https://byabbe.se/datadriven-samlingsforskning/kartor/konstnarsliv.html



Peter Paul Rubens

Hur mycket information har NM, SHM och Wikidata om denna person?

Antal påståenden per källa
Nationalmuseum
Statens historiska museer
Wikidata
Nationalmuseum
Statens historiska museer
Wikidata

Vilka andra institutioner har information om denna person?

Library of Congress
Bibliothèque nationale de France
Uppsala University (Alvin)
National Library of Sweden (Libris)
Nationalencyklopedin
Frick Art Reference Library
Nationale Thesaurus voor Auteurs
British Museum
National Library of Poland
Center of Warsaw University Library
Rijksmuseum Research Library
Biblioteca Nacional de España
Philadelphia Museum of Art
Kunstindeks Danmark
Store norske leksikon
National Library of Czech Republic
Vatican Library
Department of Prints and Drawings of the Louvre
National Gallery of Art in Washington, DC
Art UK
Unione Romana Biblioteche Scientifiche
Finnish National Gallery
Portuguese National Library
Minneapolis Institute of Art
National Library of Israel
Pinakothek
Royal Academy in London
National Portrait Gallery in London
Städel Museum
Service des Musées de France Joconde
Svenska Institutet i Rom
Museum Boijmans Van Beuningen
Art Institute of Chicago

Figure 2: *Wikidataeffekten* makes it clear how much information Wikidata contains.

birth places over their lifetimes, something that is difficult to get an appreciation of using raw data.

7. Discussion and future work

The work in the project *Usable Authorities for Data-driven Cultural Heritage Research* has provided both the two museums and WMSE with a deep understanding of what a large-scale Wikidata-focused project requires from its participants and the benefits it can bring. These insights will be of value to any other GLAM institution interested in incorporating Wikidata in their strategic work with digital cultural heritage. Furthermore, we see the linking of the data with Wikidata as the first step to more advanced usages, where the data is utilized to create added value.

As noted previously, cleaning up and reconciling the data required a lot of manual work from the team members, work that WMSE could not assist with due to a lack of domain expertise. This should be taken into account by a GLAM that wants to embark on a similar project. This is particularly relevant when the dataset in question spans over larger subject areas; it might be necessary to involve more staff than initially planned who can contribute with their specific knowledge. Alternatively, limiting the scope of the project more tightly from the beginning

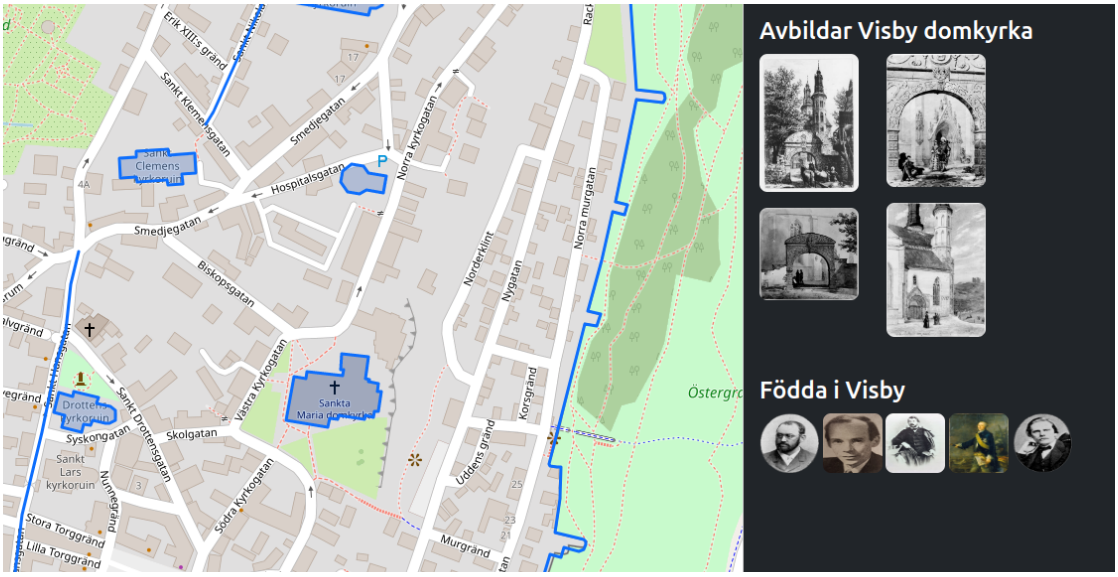


Figure 3: *Visbykartan* makes it possible to find people and artworks linked to the city of Visby.

might be considered.

As we have shown with our visualizations, Wikidata being machine-readable opens up possibilities for making cultural heritage data available to wider audiences in an attractive and interactive way. Visualization of historic data can be part of broader digital storytelling projects, where the connections between the objects and the people in museum databases are used to construct narratives that help the audience develop a more nuanced understanding of historical events and processes.[14] Here, the benefit of using Wikidata over limiting oneself to the collections of a particular institution is that much more data is available; at the same time, it poses the challenge of curating what is both relevant and trusted so as not to overwhelm the audience and ensure they are presented with accurate information.

Once an institution has ingested the Wikidata ID's of the objects or people in its collections into its own database, it makes it possible to *enrich* their own data with information from Wikidata. This is of particular interest to smaller or less resourced institutions, as it spares them the work of populating, for example, their personal authority entries with biographical information. It also makes it possible to share data that the particular institution does not find crucial to research and manage, but that can still be of interest to the users. Apart from factual information, the additional data could include links to Wikipedia articles about the topic or relevant images from Wikimedia Commons.

In practice, implementing this will be a carefully thought-out decision, as there is an understable hesitation towards using data from user-generated sources like the Wikimedia platforms. The data from Wikidata could be clearly marked as coming from an external source both in the institution's database and on its public website. In this way, the users will both benefit from the additional information and be able to decide for themselves to what degree it should be

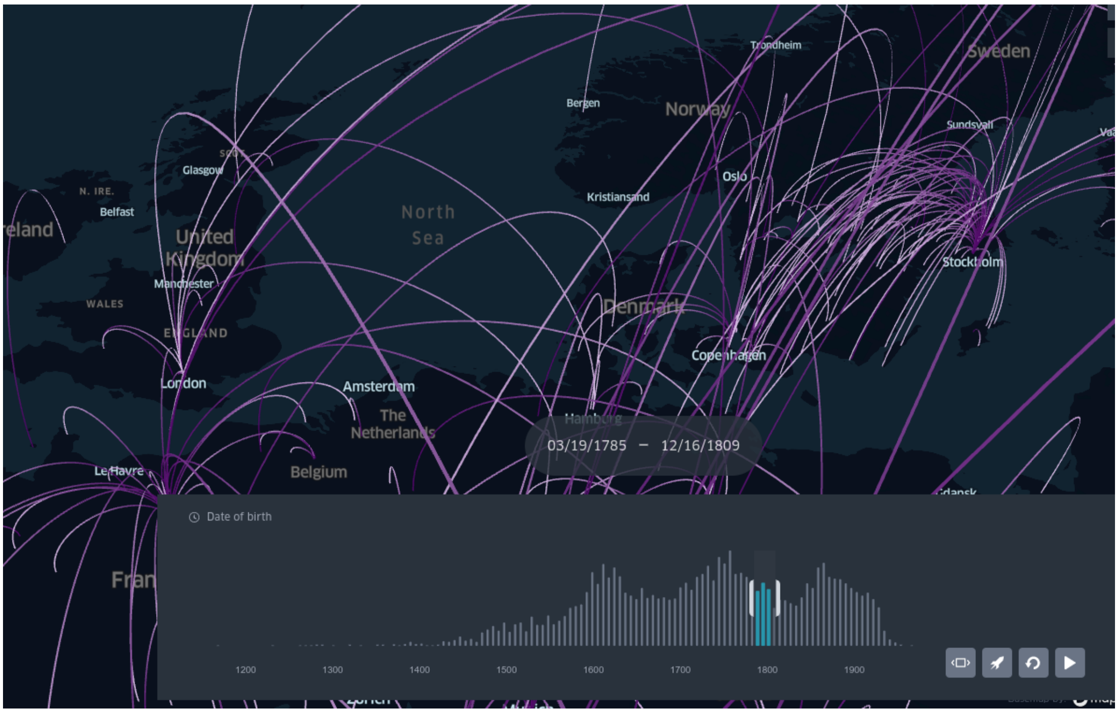


Figure 4: In *Konstnärsliv*, the user can get an overview of how artists moved between their place of birth and their place of death.

trusted. Wikidata's role as a Linked Open Data hub should also be taken into account – the other external identifiers available in the items open the door to other trusted sources, such as other cultural heritage institutions' databases. For example, it might be interesting to indicate which other museums have works by a particular artist in their collections.

Another aspect of Wikidata as a user-created database with an international contributor base is that the way it categorizes information does not have to overlap with the established practices of some scholarly communities. Periodization is a typical example; defining a period like the Middle Ages might require more nuance than a database with a simple structure is able to convey.[15] One should keep in mind that the selection and the level of detail of the data in Wikidata are a product of its, predominantly English-speaking, editor base; when establishing consensus, those who can make their voices heard are those who are shaping the information, and history and cultural heritage are particularly vulnerable to Western-centric biases.[16]

While it has to be examined with a critical eye, the third-party data in Wikidata can benefit a museum through *data roundtripping*. This is a process in which a GLAM uploads data to the Wikimedia platforms, which are then enriched by the volunteer community and the institution ingests this improved information back into its system. The Swedish National Heritage Board has researched the potential of data roundtripping in the context of the metadata of images on Wikimedia Commons, showing that while GLAM institutions have an interest in taking advantage of data generated by Wikimedians, there's both a technical barrier – extracting the modified data from Wikimedia Commons and ingesting it back – and a hesitation about the reliability of the third-party information.[17] The Biodiversity Heritage Library has successfully implemented a project where additional persistent identifiers were round-tripped from the Wikidata items of the authors represented in the collections.[18] Most recently, the Nordic Museum has announced the launch of a project in which the Wikimedia community will be invited to improve the descriptions of the newly digitized photographs from their collections, and the enriched information will then be ingested by the museum.[19]

Linking museums' collections to an LOD hub fosters collaboration and knowledge exchange between institutions, which is something that we have witnessed in the process of running the project. That the two institutions participated in it together allowed them to exchange experiences, despite the fact that both their collections and technical infrastructure differ to a great extent. The development of in-house Wikimedia and Wikidata expertise at the institutions is a net positive, as it means they are not dependent on the help of WMSE to continue working with the Wikimedia platforms after the project has ended. They have had natural opportunities to reflect on areas where their collections overlap and discuss similarities and differences in how they model their respective data. As a result, they are now able to educate their colleagues, encouraging them to e.g. add Wikidata links as part of the process of creating new authority files, and they also act as Wikimedia ambassadors in the GLAM sector, both in Sweden and internationally. From the point of view of WMSE, this is immensely valuable: in a more traditional collaborative project, where WMSE does most or all of the practical work on the Wikimedia platforms within a limited time-frame, there is a risk that the outcomes never get evaluated or implemented in the institution's digital strategy. Allowing GLAM professionals to learn and do the work themselves makes it more probable that they will feel ownership of the project and continue building on its foundations.

The team members have emphasized that the assistance of WMSE was immensely helpful when learning to navigate Wikidata and the new technical tools. GLAMs who need access to such support have several options available to them: either work with a local Wikimedia chapter, such as in the case of this project, or to bring in a so-called Wikimedian in Residence, i.e. an experienced Wikimedian to work in-house.[31] This is an established form of collaboration between organizations and the Wikimedia movement which in Sweden was done at the Swedish National Heritage board as early as 2012.[20] Wikimedians in residence have been used by libraries, museums and archives.

The project has built a solid foundation for continued, strategic work with Wikidata at both the Nationalmuseum and the National Historical Museums. In March 2023, 20% of the National Historical Museum's had been matched to Wikidata.[21] The Nationalmuseum is currently taking Wikidata into account when planning an overhaul of their database system and public website.[22] The goal is to be able to display links to the Wikidata items of the artists and their works, similar to what the National Historical Museums are doing. Preparations are also being made to work with the Iconclass system,[32] as Wikidata includes the properties Iconclass

[31] Many international examples can be found in: "Wikimedian in Residence". Meta-Wiki. Published December 22, 2022. Accessed March 17, 2023. https://meta.wikimedia.org/w/index.php?title=Wikimedian_in_residence&oldid=24275309

[32] https://iconclass.org/

notation (P1256)[33] and Depicts Iconclass notation (P1257).[34] This has the potential to further nuance the presentation of the works in the museums' collections on Wikidata.

For a GLAM institution, working with Wikidata – sharing the data, taking advantage of the work of the global community to gain new insights into their areas of work and incorporating the data from Wikidata in their own work – can seem like an interesting project, but as we have shown, there are many things to think about before embarking on this journey. The question of copyright is one that deserves to be highlighted. Importing published datasets is an established way of improving the coverage and quality of Wikidata – it would have been impossible to create 100 million items without it – but the prerequisite is that the added data has to be free from copyright restrictions.[35] Some major cultural heritage actors that have released their data under the CC0 license include the British Library in 2010[23] and Europeana in 2012.[24]

7.1. Looking into the future

One of the goals of the project has been building up both digital competence and literacy among the participants, equipping them with the means to continue working independently and share their knowledge with their peers. While digital competence encompasses the skills to use the available tools, digital literacy implies the ability to reflect and evaluate one's work with digital tools and processes, not only in relation to the needs of the institution itself, its vision and objectives, but also in a broader context – taking into account other institutions, research and society at large. Digitally literate professionals can effectively manage digital collections, engage with digital audiences, and explore innovative ways to present and interpret cultural heritage. Moving away from simply using the tools towards developing and fostering digitally literate leadership is of critical importance when making long-term plans for institutions in the GLAM sector[25] Capacity building has been pointed out by Europeana as one of the critical factors for enabling digital change, and has a prominent place in their 2020–2025 strategy.[1]

We have had this in mind throughout the project, encouraging the participants to critically evaluate how they are currently using digital tools and platforms and to envision new applications for the skills they have gained. After the project has been concluded, they will be able to make informed, strategic decisions for their respective institutions' digital development, based on a framework of openness and sharing. They can act as agents of change, a role that is crucial in institutions who wish to develop their digital capacity. [25]

Acknowledgments

The project *Usable Authorities for Data-driven Cultural Heritage Research* is financed by the Swedish National Heritage Board.

The author would like to extend their sincere appreciation to their project team at the Nationalmuseum and the National Historical Museums of Sweden, to Albin Larsson for his

[33] https://www.wikidata.org/wiki/Property:P1256

[34] https://www.wikidata.org/wiki/Property:P1257

[35] "Wikidata:Licensing". Wikidata. Published January 26, 2023. Accessed March 18, 2023. https://www.wikidata.org/w/index.php?title=Wikidata:Licensing&oldid=1819658716

work with the visualizations, as well as to their colleague at Wikimedia Sverige, André Costa, for his input on this paper.

References

[1] European Commission and Directorate-General for Communications Networks, Content and Technology, Europeana strategy 2020-2025 : empowering digital change, Publications Office, 2020. doi:doi/10.2759/524581.

[2] Decision (eu) 2022/2481 of the european parliament and of the council of 14 december 2022 establishing the digital decade policy programme 2030 (text with eea relevance), Official Journal of the European Union (2022-12-19) 4–26. URL: https://eur-lex.europa.eu/eli/dec/2022/2481/oj.

[3] F. Erxleben, M. Günther, M. Krötzsch, J. Mendez, D. Vrandečić, Introducing wikidata to the linked data web, in: P. Mika, T. Tudorache, A. Bernstein, C. Welty, C. Knoblock, D. Vrandečić, P. Groth, N. Noy, K. Janowicz, C. Goble (Eds.), The Semantic Web – ISWC 2014, Springer International Publishing, Cham, 2014, pp. 50–65.

[4] Europeana, Report on the results of the Wikimedia Taskforce, Technical Report, Europeana, 2015. URL: https://pro.europeana.eu/files/Europeana_Professional/Europeana_Network/europeana_wikimedia_taskforce_report_2015.pdf.

[5] F. Zhao, A systematic review of Wikidata in Digital Humanities projects, Digital Scholarship in the Humanities (2022). doi:10.1093/llc/fqac083.

[6] J. Neubert, Wikidata as a linking hub for knowledge organization systems? integrating an authority mapping into wikidata and learning lessons for kos mappings, in: Proceedings of the 17th European Networked Knowledge Organization Systems Workshop, 2017, p. 14–25.

[7] M. D. Wilkinson, M. Dumontier, I. J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J.-W. Boiten, L. B. da Silva Santos, P. E. Bourne, J. Bouwman, A. J. Brookes, T. Clark, M. Crosas, I. Dillo, O. Dumon, S. Edmunds, C. T. Evelo, R. Finkers, A. Gonzalez-Beltran, A. J. Gray, P. Groth, C. Goble, J. S. Grethe, J. Heringa, P. A. 't Hoen, R. Hooft, T. Kuhn, R. Kok, J. Kok, S. J. Lusher, M. E. Martone, A. Mons, A. L. Packer, B. Persson, P. Rocca-Serra, M. Roos, R. van Schaik, S.-A. Sansone, E. Schultes, T. Sengstag, T. Slater, G. Strawn, M. A. Swertz, M. Thompson, J. van der Lei, E. van Mulligen, J. Velterop, A. Waagmeester, P. Wittenburg, K. Wolstencroft, J. Zhao, B. Mons, The fair guiding principles for scientific data management and stewardship, Scientific Data 3 (2016) 160018. URL: https://doi.org/10.1038/sdata.2016.18. doi:10.1038/sdata.2016.18.

[8] A. Piscopo, E. Simperl, Who models the world? collaborative ontology creation and user roles in wikidata, Proc. ACM Hum.-Comput. Interact. 2 (2018). URL: https://doi.org/10.1145/3274410. doi:10.1145/3274410.

[9] V. Alexiev, P. Tarkalanov, N. Georgiev, L. Pavlova, Bulgarian icons in wikidata and edm, Digital Presentation and Preservation of Cultural and Scientific Heritage 10 (2020) 45–64. URL: https://dipp.math.bas.bg/dipp/article/view/dipp.2020.10.2. doi:10.55630/dipp.2020.10.2.

[10] K. M. Hill, In search of useful collection metadata: Using openrefine to create accurate,

complete, and clean title-level collection information, Serials Review 42 (2016) 222–228. doi:10.1080/00987913.2016.1214529.

[11] C. M. Strickler, Mind the wikidata gap, Atla Summary of Proceedings (2021) 301–314. doi:10.31046/proceedings.2021.2978.

[12] G. Garcia, L. Gant, Getty vocabularies openrefine tutorial, 2020. URL: https://www.getty.edu/research/tools/vocabularies/obtain/getty_vocabularies_openrefine_tutorial.pdf.

[13] D. Goldfarb, D. Merkl, Visualizing art historical developments using the getty ulan, wikipedia and wikidata, in: 2018 22nd International Conference Information Visualisation (IV), 2018, pp. 459–466. doi:10.1109/iV.2018.00086.

[14] S. Boyd Davis, O. Vane, F. Kräutli, Using data visualisation to tell stories about collections, in: Electronic Workshops in Computing, BCS Learning & Development, 2016. doi:10.14236/ewic/eva2016.44.

[15] A. Ambrosiani, M Larsson, Hur FAIR är svensk digitiserad kulturarvsdata idag?, Lychnos: Årsbok för idé- och lärdomshistoria (2023). doi:10.48202/24081.

[16] S. Cook, The uses of wikidata for galleries, libraries, archives and museums and its place in the digital humanities, Comma 2017 (2019) 117–124. doi:10.3828/comma.2017.2.12.

[17] M. Zeinstra, Returning commons community metadata additions and corrections to their source, Technical Report, Swedish National Heritage Board, 2019. URL: https://upload.wikimedia.org/wikipedia/commons/9/91/Research_Report_%E2%80%93_Returning_commons_community_metadata_additions_and_corrections_to_source.pdf.

[18] J. Dearborn, S. Leachman, Bhl is round tripping persistent identifiers with the wikidata query service, Blog Post, 2023. URL: https://blog.biodiversitylibrary.org/2023/02/round-tripping-persistent-identifiers-with-wikidata-query-service.html.

[19] Nordic Museum, 100 000 bildminnen, Press Release, 2023. URL: https://www.nordiskamuseet.se/artiklar/100-000-bildminnen.

[20] J. Carlström, L. Guldbrandsson, A. Costa, Wikipedian in Residence på Riksantikvarieämbetet, Technical Report, Swedish National Heritage Board, 2013. URL: https://www.diva-portal.org/smash/record.jsf?pid=diva2:1234493.

[21] L. Lundin, Blog Post, 2023. URL: https://shm.se/blog-article/20-av-auktoritetsposterna-kopplade-till-wikidata/.

[22] E. Eriksson, Personal Communication, 2023.

[23] J. Park, The british library releases 3 million bibliographic records into the public domain using cc0, 2010. URL: https://creativecommons.org/2010/11/22/the-british-library-releases-3-million-bibliographic-records-into-the-public-domain-using-cc0/.

[24] Europeana, Europeana opens up full dataset for re-use, Blog Post, 2012. URL: https://pro.europeana.eu/post/europeana-opens-up-full-dataset-for-re-use.

[25] J. Finnis, A. Kennedy, The Digital Transformation Agenda and GLAMs: A Quick Scan Report for Europeana, Technical Report, 2020. URL: https://digipathways.co.uk/resources/the-digital-transformation-agenda-and-glams/.

Storage over Rendition. Call for a Sustainable Infrastructure in the Digital Textual Heritage Sector with a Particular Interest in Digital Scholarly Editions

Katrine F. Baunvig[1,2,3], Krista S. G. Rasmussen[1,3], Kirsten Vad[1,3] and
Per Møldrup-Dalum[2,3]

[1] *Center for Grundtvig Studies (Aarhus University), Jens Chr. Skous Vej 3, 8000 Aarhus C, Denmark*
[2] *Center for Humanities Computing, Jens Chr. Skous Vej 4, Building 1483, 4th floor, 8000 Aarhus C, Denmark*
[3] *Aarhus University, Nordre Ringgade 1, 8000 Aarhus C, Denmark*

Abstract

A significant amount of human and pecuniary resources has gone into the production of the long line of digital scholarly editions that within recent decades have sprung to life in Scandinavia, in the Baltic region, as well as in the rest of Europe. Notwithstanding the heritage perspective, the sector is paradoxically characterized by a presentist preoccupation with instant results – first and foremost with the rendition of the given data set. Concerns for long-term data management perspectives – that is: interest in the post-production afterlife of the data – is relatively meagre. This goes for project managers, for project host institutions, and for the research foundations financially supporting the projects. So, despite harmonizing initiatives at production level and pre-edition compilation initiatives, such incentives promote a situation of insulated digital scholarly editions focusing on unique URLs and distinctive qualities of the given material. This hinders project synergy in the production phase. Moreover, it hinders the construction of long-term and sustainable data management solutions. To remedy this situation, we propose a clear division of labour between the tasks of data production, of data rendering, and of data storing. This division should ideally be sought at an institutional level. This will secure the accumulation of know-how in teams refining the respective workflows. In addition, we encourage private and public foundations to bolster this infrastructure by making project compliance a criterion for funding.

Keywords

Scholarly Editions, Digital Scholarly Editions, Cultural Heritage Sector, Data Storage, Data Preservation, Data Production, Data Rendition, Research Infrastructure

DHNB2023 | Sustainability: Environment - Community - Data. The 7th Digital Humanities in the Nordic and Baltic Countries Conference. Oslo – Stavanger – Bergen, Norway. March 8–10, 2023.

baunvig@cas.au.dj (K. F. Baunvig); ksgr@cas.au.dk (K. S. G. Rasmussen); kiva@cas.au.dk (K. Vad); per@cas.au.dk (P. Møldrup-Dalum)
https://au.dk/baunvig@cas.au.dk (K. F. Baunvig); https://au.dk/ksgr@cas.au.dk (K. S. G. Rasmussen); https://au.dk/kiva@cas.au.dk (K. Vad); https://au.dk/per@cas.au.dk (P. Møldrup-Dalum)
0000-0001-6706-6694 (K. F. Baunvig); 0000-0003-3424-2334 (K. S. G. Rasmussen); 0000-0002-8300-4597 (K. Vad); 0000-0001-9976-3515 (P. Møldrup-Dalum)

DHNB Publications, DHNB2023 Conference Proceedings, https://journals.uio.no/dhnbpub/issue/view/875

1. Introduction

This paper presents little short of an infrastructural vision for the digital cultural heritage sector with particular interest in the area of digital scholarly editions [DSE]. We hope to convince stakeholders in the Nordic region that it is high time for a division of labor within the field. That is, we hope to persuade fellow project managers, project host institutions, GLAM personnel, politicians, and the funding agencies supporting the projects, to join forces and aim for two goals.

1) First, we have to break the main DSE tasks into three (relatively) individual procedures: a. The first concerning to the pre-production, production, and post-production of a given material. b. The next regarding the rendition of this material to given current target groups. c. The third pertaining to the secure and long-term storage of the data material.

2) Meeting UNESCO's call for open science and a sustainable research data infrastructure, we aim to coordinate, streamline, and bolster the data storage task at a regional level – in a FAIR manner. The FAIR principles for scientific data management and stewardship seek to secure the Findability, Accessibility, Interoperability, and Reuse of digital assets.

The need for moving the field in this direction is both internally and externally conditioned. Internally: the DSEs have within the last couple of decades accumulated a significant amount of general know-how and routine that we, due to the somewhat fragmented nature of the project landscape dominated by a silo-mentality, still have not been able to reap the potential synergy rewards from. Externally: the push for an open science infrastructure and for FAIR data management plans has intensified within the last five years. In other words, the time for action is thus now: On the one hand the DSEs have reached maturity as a scholarly field and are now able to obligate cross-project coordination, on the other hand, data management procedures have improved radically while the data management impetus and ethos have intensified correspondingly.

In sum: Notwithstanding the heritage perspective, the sector is in other words paradoxically characterized by a presentist preoccupation with instant results – first and foremost with the rendition of the given data set. There seem to be somewhat standardized operations in place for the production and rendition of digital scholarly editions, however, solutions for long-term data management perspectives – that is: the postproduction afterlife of the data – are as of yet unconsolidated. This is the situation among project managers, among project host institutions, and among the research foundations financially supporting the projects. So, despite harmonizing initiatives at production level and pre-edition compilation initiatives (e.g., www.litteraturbanken.se), the incentive structure still promotes a situation of insulated digital scholarly editions focusing on unique URLs and distinctive qualities of the given material. This hinders project synergy in the production phase. Moreover, it hinders the construction of long-term and sustainable data management solutions.

2. The History of Scholarly Editions

Scholarly editions of text corpora of various natures have a long and prolific history. Critical Bible studies and studies of the works of the philosophers of Roman and Greek Antiquity –

established as a scholarly practice in the Renaissance, refined in the Early Modern Reformation period, and accelerated throughout the Enlightenment Era – are the deep roots of the endeavors. Through the course of the 19th century, the procedures of preserving and rendering given sets of text as close to the originals as possible were, however, broadened to include phenomena in vogue: the current 'Genius' and the 'Hero'. In other words: the authorships or oeuvres of great men whose lives and deeds were still fresh in collective memory. This was the period of the emerging nation-states' construction of cultural identities and formation of a cultural heritage canon; scholarly editions aided in this process and as a result scholarly editing gradually grew into a field proper (p. 231-266) [1]. Through the 20th Century, the field consolidated into two main areas: classical philology, concerned with codices, and new philology centered on printed books. [1]. Within the Nordic countries, these two fields operate somewhat independently.

3. The History of Digital Scholarly Editions

Scholarly editing has always been pivotal for the development of the so-called Digital Humanities, whose origin is often dated to the year 1949 when Roberto Busa, in collaboration with IBM, began working on his concordance of the works of *Thomas Aguinas: Index Thomasticus.* Busa's work is widely recognized as the birth of the field 'humanities computing', although his index is not a scholarly edition per se. *Index Thomasticus* was first published in the 1970s and is now available online.[2]. The first digitization of a canonical Nordic author also formed the basis of an index when the Canadian philosopher Alastair McKinnon, affiliated to McGill University in Montreal, had all of Søren Kierkegaard's writings transferred on to punched cards in the 1960s. The indices were published in print, and do not now exist digitally.[3] This digitalization later morphed into a digital edition, which was among one of the first in the world. This edition again morphed in to the latest scholarly edition of Kierkegaard's oeuvre: *Søren Kierkegaards Skrifter (The Writings of Søren Kierkegaard)* .[4]

These editions mirror the general development of digital scholarly editions: There is a distinct pre-internet era, in which editions were digitally prepared, but mainly distributed in print (or in some cases CD-ROM). Then follows the internet era in which editions move online, and much theorization and effort was invested in creative rethinking of scholarly editions. The field has consolidated, and the present-day situation is more or less post-internet and occupied with the redefinition of digital scholarly editions that are both sustainable and viable on the long haul of the post-production stage.

[1] New philology in this sense only designates the material, other uses of the term 'new philology' are material printed (within medieval studies), or focused on native-language sources within indigenous societies. See: (p. 999-1006) [2], [3]

[2] http://www.corpusthomisticum.org/it/index.age

[3] Alastair McKinnon: The Kierkegaard Indices. Vol. 1–4. Leiden 1970–1975.

[4] Rasmussen, Krista Stinne Greve: 'Bytes, Books and Readers: Some Clues to the History of Alastair McKinnon's Digital Edition of Søren Kierkegaard's Samlede Værker', Editio: Internationales Jahrbuch für Editionswissenschaft, 30:1, 2016, pp. 184-196. https://doi.org/10.1515/editio-2016-0012

4. Digital Scholarly Editions: Challenges to the Sustainability of the Field

For a long period of consolidation, the DSE have, thus, as a professional community built up and refined DSE routines. It is an obvious fact but not a trivial observation that a significant amount of cultural heritage data has come to life in digital formats within the last decades. It is also an obvious fact but not a trivial observation that a significant amount of human and pecuniary resources has gone into the production of the long line of DSEs that within recent decades have sprung to life in Scandinavia, in the Baltics, as well as in the rest of Europe. The projects piling up on the Berlin-based Institut für Dokumentologie und Editorik's "A catalogue of Digital Scholarly Editions" (https://v3.digitale-edition.de/vlet-collected-works.html) testify to an overall European trend pertaining to resource allocation. The digitization and computational exploration of cultural textual heritage material attract funding (Rasmussen et al. 2022). This we take as a sign of maturation within the field. Two trends, however, clearly pose a hindrance or a challenge to the sustainability of DSE as a mature field. One is the centripetal power of the canon, the other is the dominating silo-mentality.

4.1. The Centripetal Power of the Canon: The Case of the DSE Grundtvig's Works

In Danish public discourse poet, pastor, politician, and romanticist N.F.S. Grundtvig (1783–1872) is regarded as one of the most (if not the) central figure in the nineteenth-century Danish nation building process, as well as in the construction of a modern Danish Christianity: In short, he is regarded as a cultural saint.[5] In scholarly literature, it is widely acknowledged that Grundtvig sought to stimulate the process of assembling a collective Danish emotional consciousness based on 1) a horizontal-contemporary axis incorporating the different strata within the socially heterogeneous "Folk" and on 2) a vertical-historical axis connecting present-day Danes with forefathers and legendary characters. In social historian Benedict Anderson's words, the emotional fabric intended by this attempted interlacing was an imagined community. Nowadays, Grundtvig's cultural imprints are acknowledged by most Danes: "N.F.S. Grundtvig founded Danish democracy"; "N.F.S. Grundtvig established the Church of Denmark (folkekirken)"; "N.F.S. Grundtvig is the founder of the Danish school system"; "N.F.S. Grundtvig revived the pre-Christian Nordic tradition"; "N.F.S. Grundtvig is the most important writer of Christian hymns in Denmark". These are surprisingly recurrent statements in Danish public media, deeming his intellectual activity more culturally important than the work of his world-famous contemporaries Søren Kierkegaard (1813–1855) and Hans Christian Andersen (1805–1875).

Considering Grundtvig's cultural status, it is perhaps no surprise that a consortium including members of the Danish parliament some 15 years ago decided to pave the way for the creation

[5] This 'sainthood' is not a banal, cosmetic analogy. Reverence and quasi-ritual structures have been built around Grundtvig as a 'Great Dead' [4]. Grundtvig has a cathedral named after him: the Copenhagen Grundtvigs Kirke; every year his birthday (almost coinciding with the day of his death) is celebrated in Grundtvig-relevant institutions; one such celebration entails the opening of his crypt at the small cemetery Clara's Kirkegård on the outskirts of the Sealandic town of Køge. Moreover, Grundtvig's Death (Grundtvigs død) is a commodity – at least it is a recent title in a popular book series by Aarhus University Press written by Grundtvig scholar Jes Fabricius Møller (2019).

of a scholarly edited, digital version of Grundtvig's published writings, making them available to Danish citizens free of charge. The writings, published within a period of 68 years (1804–1872), amount to 4M word tokens distributed over approximately 37K standard pages.

First editions of each work are OCR prepared and are now being manually cleansed. A crucial step in producing an accurate digital corpus is this labor-intensive cleansing and annotation of the raw and oftentimes somewhat dirty OCR results. This corpus forms the basis of the digital scholarly edition produced by a group of ten Grundtvig-specialized scholarly editors – philologists trained in fields relevant to the domestication of Grundtvig's prose, such as (obviously) nineteenth-century Danish, but also Old Norse, Greek, Bible Studies, hymnology, romantic philosophy, eighteenth-century historiography, political history, etc. As part of the edition, this équipe furnishes the individual texts with contextualizing introductions and glossaries. Their work, piling up on www.grundtvigsvaerker.dk, was initially estimated to consume 200 man-years. This prognosis seems to hold: 11 years and 100M DKK (approx. 15M Euro) later, the ten scholarly editors are halfway through the project. Such details are highly relevant when drawing up the contours of DSE – and to some extent also of the Humanities Computing – because the thorough markup is what leaves the data open to the general public as well as to comprehensive, fine-grained, hermeneutically complex scholarly explorations.

Clean, reliable, and flexible but the point here is that it is also highly exclusive. High quality data is, as we have just tried to spell out, burdensome to create in terms of time and funding. AND not every type of material, not every type of authorship will move politicians and research foundations to cover the expenses[6] - for the last decades mainly canonical figures and archives have done so. A long line of political and ethical problems nest here.

But gradually funding of corpora of a different nature seems to circumvent this drift towards the canon – or the canon seems to evolve in a more inclusive, socially sustainable direction. If this drift catches on, it would be a great gain for the cultural heritage sector, for humanist scholars, and for the general public.

4.2. The Silo-Mentality dominating the DSE Sector

The early and consolidating days of the DSE sector have been characterized by a long line of individual projects with no or not that much infrastructural coordination. Obviously, personal, and scholarly networks have sought to mitigate the fact that it was every DSE project for itself / that every project ultimately was responsible for each stage in the DSE production line – this fostered what could perhaps be thought of as a silo mentality. Highly different in nature and scope, NNE (Nordisk Netværk for Editionsfilologer) and TEI are two obvious examples of initiatives aiming to enhance coordination among individual projects. Though much great work has been done in, around, and with NNE and TEI, we see them as symptoms of a core problem that there is a lack of harmonization, coordination, or division of labor within the sector. One obvious reason being that each edition is considered a full-scale research project in itself, mimicking scholarly editions of the print era. However, a print and a digital edition are not the same, storage and rendition being the main game changers.

[6] The Great Unread vs. The Canon has long been a topic in Literary History (Margaret Cohen, The Sentimental Education of the Novel (Princeton: Princeton University Press, 1999), Franco Moretti, "The Slaughterhouse of Literature" Distant Reading (London: Verso, 2013)).

Many steps have been taken to move scholarly editions into the digital paradigm, which has affected the production in almost every step. Editors have learned TEI and XML, and many have even attained the expertise to produce their editions online. Now they struggle with the worries of storage, and many probably spend troublesome hours faced with a field of expertise, they–rightly–feel uneasy to enter, that is long-term preservation of digital, cultural heritage. The difference between print and digital editions is nowhere as prominent as when it comes to rendition and storage. Editors of print editions leave these concerns mostly in the hands of printing houses and libraries. This is still not the case with digital editions, and more often than not rendition and storage end up in the silo of the scholarly edition.

Fully adhering to the digital paradigm means that rendition and storage have to be non-silo endeavours. Research infrastructure and joint publishing tools have been promoted,[7] but hitherto the responsibility for and expertise into long-term rendition and storage have resided with the editor. An approach that is neither sustainable nor viable.

5. Digital Scholarly Editions: A Mature Field

Challenges aside, a cross-sectoral routine or task flow seems to have somewhat stabilized during the last decade. Every project thus experience having to deal with a given material in need of processing by way of, e.g., OCR or Transcribus. The next step running across the DSE field is the manual cleansing and mark-up of the given material carried-out by digital scholarly editors. The result hereof is what is rendered on given URLs for lay and scholarly users to consume; scholarly consumption of the DSE data can, however, also circumvent the rendition and focus alone on the 'raw' digitized material. Finally, we have the stage of storage. Here, however, no specific long-term model seems to have stabilized (see Fig. 1).

Researchers have, furthermore, consolidated the field of editorial theory, both at a Nordic and European level, and have shared forces in the constructions of highly elaborate digital scholarly editions on the internet. In this sense it is a mature field that is steering towards a new stage in the history of digital scholarly editions: the digital afterlife. Preservation of the current digital scholarly editions that resides on the internet is facing an obvious problem: they are not maintainable in their current design [5]. The steps we need to take in the immediate future are best taken together within the field that is both a mature research field and a consolidated community.

6. Conclusion: Tripartite Division of Labour

We hereby invite stakeholders involved in digital scholarly editing to remedy this situation. We propose to seek binding regional infrastructures articulating and dividing responsibility for a) the production of the data, b) the short-term rendition of the given data sets, and c) for the long-term storage of data in a FAIR manner. The underlying logic is that data storing represents a humdrum operational task with few rewards in terms of potential institutional exposure and public acknowledgement. This explains the slapdash and unambitious solutions available.

[7] Such as i.e. TEI Publisher and CLARIN.

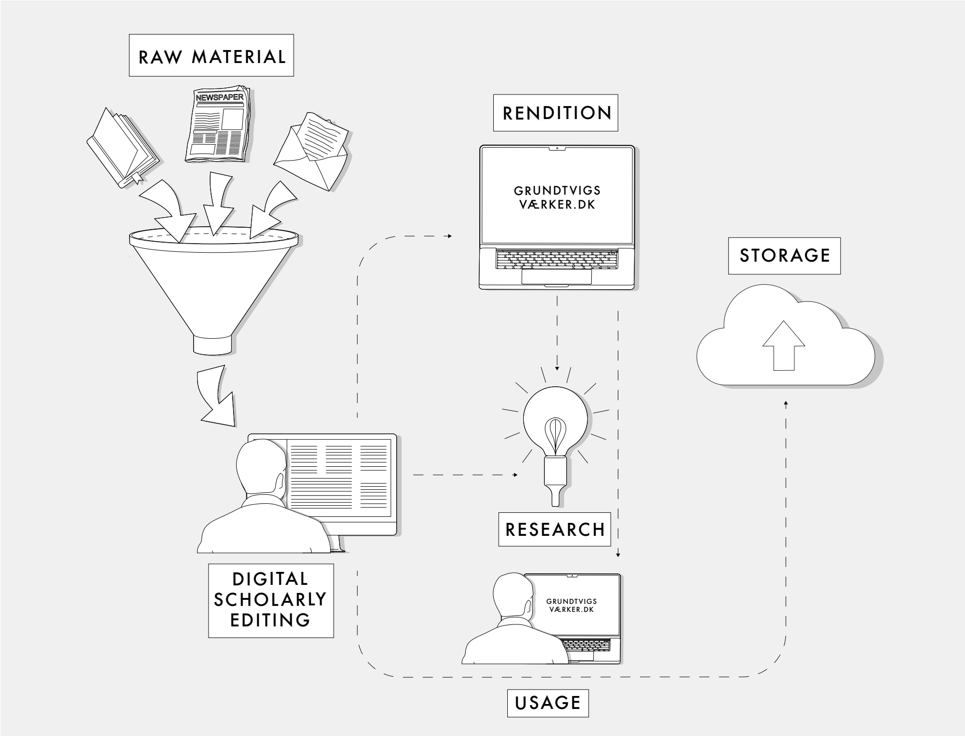


Figure 1: The DSE Work Loop. Every project will experience having to deal with a given material in need of processing by way of, e.g., OCR or Transcribus. The next step is the manual cleansing and mark-up of the given material carried-out by digital scholarly editors. The result hereof is what is rendered on given URLs for lay and scholarly users to consume; scholarly consumption of the DSE data can, however, also circumvent the rendition and focus alone on the 'raw' digitized material. Finally, we have the stage of storage. Here no specific long-term model seems to have stabilized.

Nevertheless, proper storing is the only sustainable argument for the resources going into the production of the digital editions.

In conclusion: We propose a clear division of labour between the tasks of data production, of data rendition, and of data storage. This division should ideally be sought at an institutional and a regional level. This will secure the accumulation of know-how in teams refining the respective workflows. In addition, we encourage private and public foundations to undergird this infrastructure by making project compliance a criterion for funding.

References

[1] J. Kondrup, Dansk Editionshistorie, bd. 1-4, Museum Tusculanum, 2021.
[2] S. Yager, New Philology, in: New Philology, De Gruyter, 2010, pp. 999–1006.

URL: https://www.degruyter.com/document/doi/10.1515/9783110215588.999/html?lang=en. doi:10.1515/9783110215588.999.

[3] M. Restall, A History of the New Philology and the New Philology in History, Latin American Research Review 38 (2003) 113–134. URL: https://www.jstor.org/stable/1555436, publisher: Latin American Studies Association.

[4] R. Bartlett, Why Can the Dead Do Such Great Things?: Saints and Worshippers from the Martyrs to the Reformation, Princeton University Press, Princeton, NJ, 2013.

[5] E. Oltmanns, T. Hasler, W. Peters-Kottig, H.-G. Kuper, Different Preservation Levels: The Case of Scholarly Digital Editions, Data Science Journal 18 (2019) 51. URL: http://datascience.codata.org/articles/10.5334/dsj-2019-051/. doi:10.5334/dsj-2019-051, number: 1 Publisher: Ubiquity Press.

The Laborious Cleaning: Acquiring and Transforming 19th-Century Epistolary Metadata

Senka Drobac[1], Johanna Enqvist[3,2], Petri Leskinen[2,1], Muhammad Faiz Wahjoe[1], Heikki Rantala[1], Mikko Koho[1], Ilona Pikkanen[3], Iida Jauhiainen[3], Jouni Tuominen[2,1], Hanna-Leena Paloposki[3], Matti La Mela[2,5] and Eero Hyvönen[1,2]

[1] *Aalto University (Semantic Computing Research Group (SeCo)), Finland*
[2] *University of Helsinki (HSSH, HELDIG, Cultural heritage studies), Finland*
[3] *The Finnish Literature Society, Finland*
[5] *Uppsala University, Sweden*

Abstract

The paper documents the process of collecting, consolidating, and publishing epistolary metadata from Finnish cultural heritage organizations to create an archive for bottom-up analyses of 19th-century epistolary culture. We describe and discuss the data survey that was conducted to gather information about available letter collections across Finland, as well as the cleaning and harmonizing of over 350,000 letters from twelve different sources in various digital formats. We have also developed a data model that combines event-based and letter-based aspects of the metadata. Furthermore, the paper contributes to the ongoing discussion of the initial phases of data-intensive research and the importance of discussing the labor of cleaning data. We believe that our experiences described in this paper can have wider significance for other digital humanities projects in Europe.

Keywords

Letter Metadata, Semantic web, Linked Open Data, 19-th Century

1. Introduction

This paper describes the process of gathering, aggregating, harmonizing, and publishing epistolary metadata from Finnish cultural heritage (CH) organizations in order to create an *inclusive* archive for bottom-up analyses of 19th-century epistolary culture in the Grand Duchy of Finland (1808/09-1917). The authors are working in the digital humanities consortium project *Constellations of Correspondence (CoCo)* [1] that aggregates and publishes 19th-century epistolary metadata from scattered collections of Finnish CH organizations. The unified collections are harmonized, linked, enriched, and published on a Linked Open Data (LOD) service, and as a semantic web portal.

DHNB2023 | Sustainability: Environment - Community - Data. The 7th Digital Humanities in the Nordic and Baltic Countries Conference. Oslo – Stavanger – Bergen, Norway. March 8–10, 2023.

senka.drobac@aalto.fi (S. Drobac)
0000-0002-7645-3079 (S. Drobac); 0000-0003-0901-7987 (J. Enqvist); 0000-0003-2327-6942 (P. Leskinen); 0000-0002-4716-6564 (H. Rantala); 0000-0002-7373-9338 (M. Koho); 0000-0001-9435-7163 (I. Pikkanen); 0000-0003-4789-5676 (J. Tuominen); 0000-0003-1412-8622 (H. Paloposki); 0000-0003-0340-9269 (M. La Mela); 0000-0003-1695-5840 (E. Hyvönen)

DHNB Publications, DHNB2023 Conference Proceedings, https://journals.uio.no/dhnbpub/issue/view/875

In what follows we will scrutinize the different phases of the data acquisition and processing. First, we will discuss a data survey that was sent to a wide variety of Finnish CH organizations in order to acquire systematic and comparable information as to their collections. Second, we will describe the stages of processing and cleaning the epistolary metadata. We began with more than 350 000 letters, from twelve different sources, each in its own digital format. Although the received data is mostly structured, we needed to parse running text to retrieve metadata in nearly every collection. Moreover, we had to analyze each dataset and identify possible structural mistakes. Furthermore, some records required Natural Language Processing to get actor names (e.g. senders, recipients) in dictionary format. The most difficult task has been to process 400 Word files provided by the National Library of Finland, which contain correspondence metadata in a variety of formats, easily understandable to humans but difficult for computational processing.

Furthermore, we explain the efforts made to create a harmonizing data model for epistolary metadata collections that adhere to international standards. The data model is designed to support modeling of the relevant properties of letter metadata collected from source datasets, to promote interoperability, and to support efficient use of data in e.g. SPARQL queries and the semantic portal developed during the project.

We will round off the paper by discussing the initial phases of data-intensive research and how this time-consuming "data work" should be described, understood, and credited. [2]

Although this paper describes and discusses the processing of Finnish epistolary metadata (or metadata that has ended up in the Finnish archives and museums), we believe that our experiences may have wider significance. In Europe, there are several digital humanities projects that harvest well-curated metadata (detailed information about senders, recipients, dates, and places) from edited letter collections – like Europeana[1] [3], Kalliope Catalogue[2], The Catalogus Epistularum Neerlandicarum[3], Electronic Enlightenment[4], ePistolarium[5] [4], SKILLNET[6], correspSearch[7], the Mapping the Republic of Letters project[8], NorKorr - Norwegian Correspondences and Linked Open Data [5], and Early Modern Letters Online (EMLO)[9] [6, 7, 8]. Bruneau et al. discuss applying Semantic Web Technologies to modelling the correspondences of French scientist Henri Poincaré and publishing on an online portal[10] [9]. The data work described in this article can therefore serve as a precedent for future projects, that set out to acquire letter metadata from the collections of cultural heritage organisations on a wider scale.

[1] http://www.europeana.eu

[2] http://kalliope.staatsbibliothek-berlin.de

[3] http://picarta.pica.nl/DB=3.23/

[4] http://www.e-enlightenment.com

[5] http://ckcc.huygens.knaw.nl/epistolarium/

[6] https://skillnet.nl

[7] https://correspsearch.net

[8] http://republicofletters.stanford.edu

[9] http://emlo.bodleian.ox.ac.uk

[10] http://henripoincare.fr/s/correspondance/page/accueil

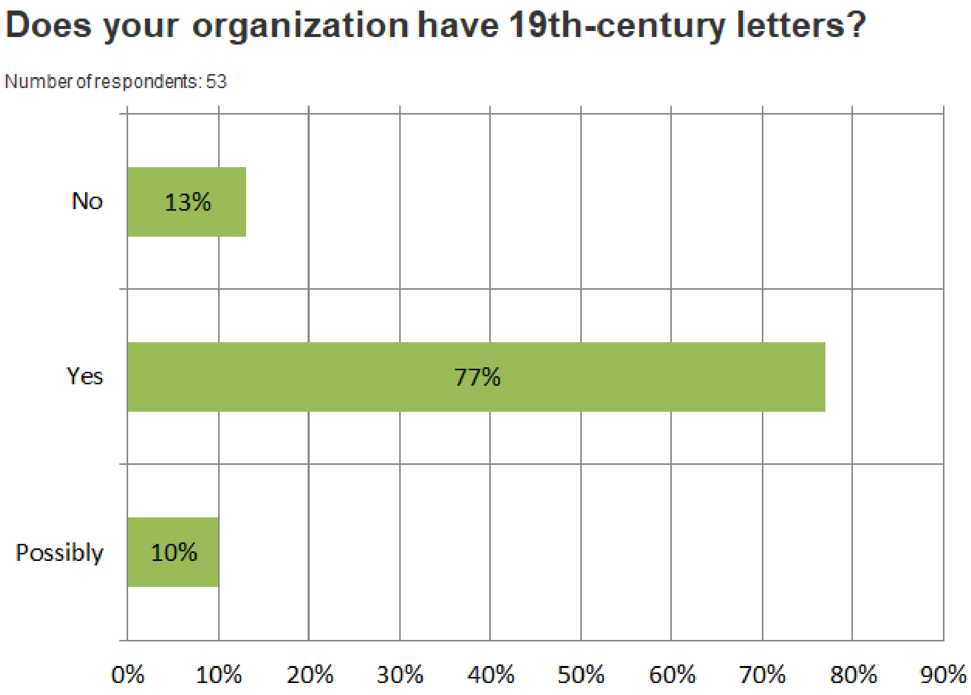


Figure 1: The Webropol survey for Finnish CH organizations: percentage of organizations governing 19th-century letters

2. Data Acquisition

In the first phase of the project, when planning the gathering of metadata from the Finnish CH organizations, we soon realized that to form an overview of the field we must try to investigate how much 19th-century correspondence there actually is in the Finnish CH organizations – archives, libraries and museums. This information was not in general pre-available from the organizations themselves; we will get to the reasons later in this article.

To get the needed information and to approach the possible governors of 19th-century letters, we composed a Webropol metadata survey with 18 questions, including both fundamental questions about the existing letter material and more detailed questions regarding the letters and their metadata in individual CH organizations. The survey stresses that in order to obtain a reliable overall picture, we would also welcome responses from those organisations that do not have 19th century letter collections.

Since the early 2022, we have sent links to the survey, and have partially re-sent them to 102 CH organizations, ranging from large institutions to small local museums and archives. By the end of February 2023, we had received 53 responses to the survey. We consider this response rate to be significant as it appears that the majority of CH organizations with large letter collections have responded. There is also a public web link available, which has been shared via blogs and

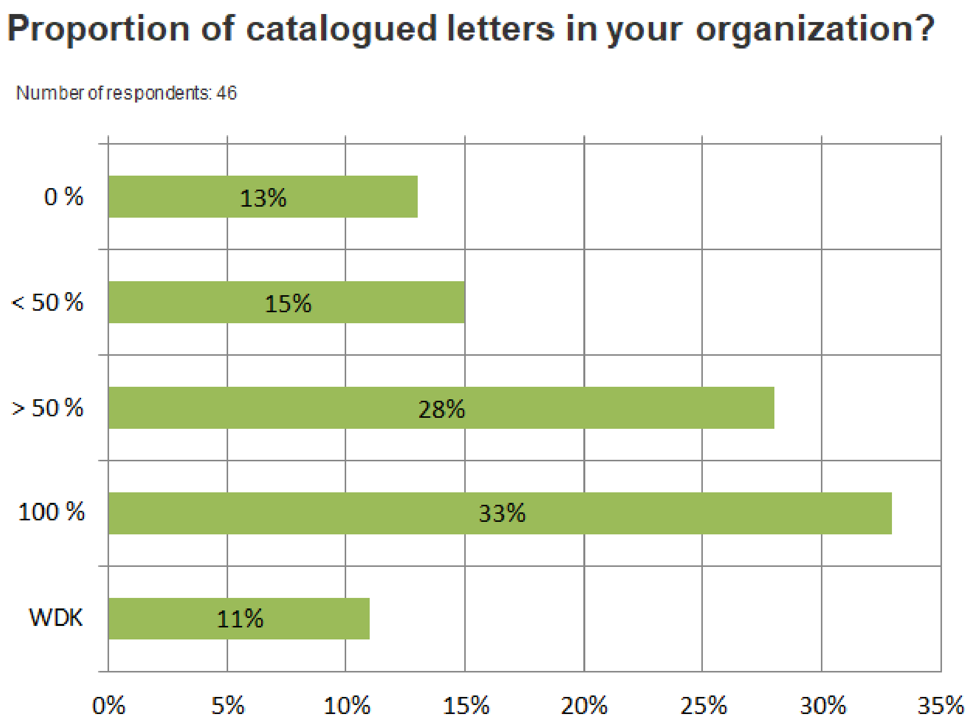


Figure 2: The Webropol survey for Finnish CH organizations: proportion of catalogued 19th-century letters

in social media. The biggest national CH organizations, the National Archives of Finland, the National Library of Finland and the Finnish National Gallery, had already in advance agreed on submitting their letter metadata and did not participate in the survey.

Once we analyzed the results of the survey, we found that 77% of respondents indicated that they have 19th-century letters in their collections, as illustrated in Figure 1. Furthermore, the majority of these organizations expressed their interest in cooperating by submitting their letter metadata. As of writing this article, we have received metadata from 12 organizations. More material is coming in and the data migration has already been agreed with five more organizations. Gathering metadata is thus an ongoing process and will continue in 2023–24.

While trying to get the idea of the overall number of preserved 19th-century letters, we have come up to the fact the CH organizations very often do not know the exact number of letters in their collections. It is maybe characteristic that 10% of the respondents answered that they possibly have 19th-century letters. The answers to the question of the percentage of catalogued letters in their collections give an explanation to this. As Figure 2 illustrates, only 33% of the respondents with these collections have catalogued all their letter material, 28% over a half of them, 13% not at all and 11% could not give estimation. This tells about the everyday realities in CH organizations: they do not have enough resources to organize and catalogue all their

Table 1
Overview of the collections being processed, including the collection name, total number of letters, and format of the data received. The first three collections are well-curated and edited, while the remaining collections are obtained from various institutions.

Name	Size (Letters)	Format
Albert Edelfelt	1 600	JSON Web API
Elias Lönnrot	6 247	JSON Web API
J.V. Snellman	1 514	RDF
Svenska litteratursällskapet i Finland (SLS)	43 000	XLS files
National Gallery	9 976	CSV
Finnish Art Society	1 147	XLSX
National Archive	295 000	CSV
Åbo Akademi	366 614	XML
The Finnish Literature Society (SKS)	37 676	XLSX
HS Foundation	2 500	CSV
Postal Museum	50	XLSX
National Library	unknown	Word files

archival collections.

Computational methods can naturally only make use of metadata in a structured form, i.e. in a form that can be processed computationally. In addition to the lack of overall quantitative data on the collections, we did not know beforehand how much of the data has been described (catalogued) by the archivists and what proportion of this metadata is computationally accessible. According to our survey, only 35% of the existing catalogues on 19th century letters are in some electronic form: in databases, as word or excel documents.

The answers to the survey reflect the fact that the letter collections have accumulated in archival collections over decades, sometimes centuries. Consequently the metadata production has deep temporal layers too. Even within a single collection, correspondence may have been described at different stages and on different platforms. Descriptive practices and concepts have evolved over time, but also in relation to the specificity of the material being described and the creative solutions of individual archivists. Today's digital systems are promoting the production of more structured descriptive information. However, it is not self-evident that the metadata fields of the different archival systems are comparable or that the subsequent extraction of data has been planned and tested when organisational databases and archival information systems have been built.

Overall, the varying practices of storing epistolary metadata challenge the computational use of it. In the next section we describe the various solutions of processing the data, depending on the format and on the varying contents and structures of the letter catalogues.

3. Data Cleaning and Transformation

An overview of the collections that have been received and are currently being processed is presented in Table 1. The table has three columns: the first column lists the names of the collections, the second column shows the total number of letters in each collection, and the last column specifies the format of the received data. The first three rows, which are highlighted in a light gray color, correspond to well-curated published edited collections. The remaining collections are obtained directly from CH institutions, some as data dumps from their internal systems and others as copies of their storage records.

To harmonize all these different datasets, we have created a fully automatic transformation pipeline, like illustrated in Figure 3. The pipeline comprises several stages, starting with the processing of each received dataset into an intermediary RDF (Resource Description Framework) format, which contains literal values. The next step involves harmonizing the data with the CoCo Data Model, which we have developed based on international standards. To enrich the data, we link the recognized actors and places with external resources. At this point, it is possible and desirable to deduplicate actor and place names, a process that we plan to develop in the future. Finally, the transformation pipeline produces a harmonized dataset of correspondence metadata, which is structured and optimized for accurate and efficient analysis.

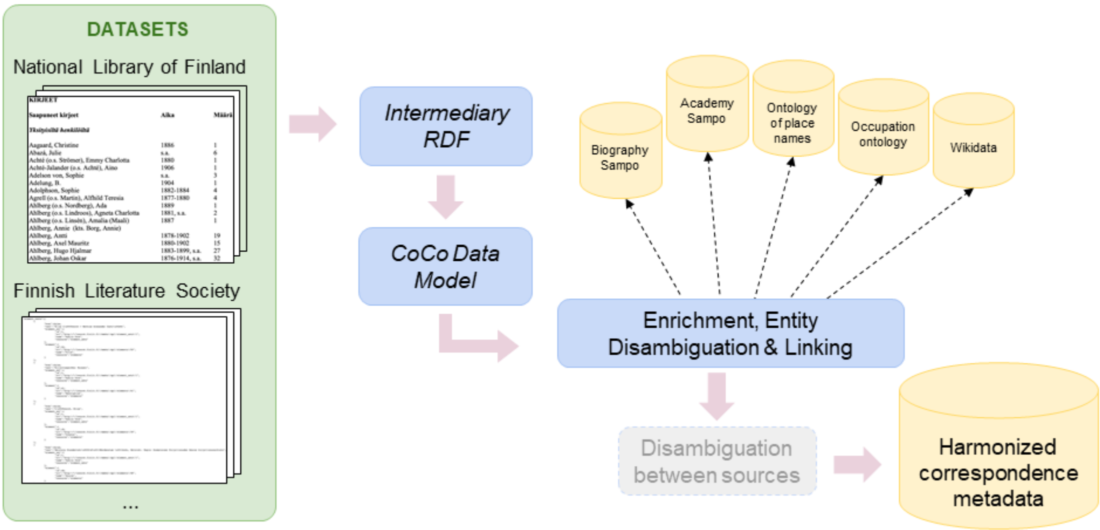

DATASETS
National Library of Finland
Finnish Literature Society
...
Intermediary RDF
CoCo Data Model
Enrichment, Entity Disambiguation & Linking
Biography Sampo
Academy Sampo
Ontology of place names
Occupation ontology
Wikidata
Disambiguation between sources
Harmonized correspondence metadata

Figure 3: Illustration of the fully automatic transformation pipeline used to harmonize heterogeneous datasets. The pipeline involves several stages, including conversion into an intermediary RDF format, harmonization with the CoCo Data Model, linking with external resources, and deduplication/disambiguation (planned for future development). The resulting harmonized dataset of correspondence metadata is structured and optimized for analysis.

Cleaning and harmonizing datasets

In the first step, we created a separate transformation process for every dataset where we locate and extract relevant information. Given the substantial variations in the nature and details of the correspondences across the datasets, we analyzed the first 11 datasets to determine a comprehensive list of extractable properties. The resulting list is presented in Table 2.

Table 2
A list of the properties we have collected from the source datasets. For easier comprehension, we can divide properties in three groups: information about sender and recipient, correspondence information, and archival information. The properties we searched for and extracted from the datasets are listed in the right column. (Not all properties are always present.)

Group of information	Collected properties
Sender and recipient	Full name, first name, last name, particle of nobility, date of birth, date of death, gender, occupation, type (e.g. person, company, family)
Correspondence information	Date of sending, amount of letters, place of sending, place of receiving, language of the letter, letter type (e.g. letter, postcard, telegram), content of the letter, translation of the letter, person reference (i.e. people mentioned in the letter), place reference (i.e. mentioned places)
Archival information	Record id, archival fonds, series/letter collection

In almost all the cases, the source datasets have information about both sender and recipient (i.e. actors). The most important thing for us is to get the full name, which can be written in different formats (e.g. *"Walleen, Carl Johan"*, *"Carolina Carlstedt"*, *"C. M. Creutz"*). Sometimes, the actor is a family (e.g. *"perhe Carlstedt"*), or institution (e.g. *"Königliche Akademie der Künste zu Berlin"*). Occasionally, there is also information about a person acting on behalf of an institution (e.g. *"Snellman, Johan Vilhelm / STY:n taloustoimikunta"*). Besides the name, additional information about the actors (such as gender and date of birth and death) is available only in few datasets, but we try to acquire it whenever available because it is important for the later disambiguation of the actors.

The available information on correspondence varies considerably across datasets. Temporal information is generally available, either in the form of a full date or just a year of sending. In most cases, however, the number of letters is reported as the total number of letters sent within certain boundary years. Occasionally, information on the place of sending is available, but information on the destination is rare. Language and letter type information is also available in some datasets, ranging from regular letters, postcards, telegrams, to less common letter types such as invitations. Information-rich datasets may even contain summaries or full contents of letters, as well as information on persons and places mentioned within them.

The archival information that we aim to capture and present includes the correspondence ID, if available, or another identifier that can pinpoint the precise location in the received dataset, such as the row ID for CSV files. We also retain the name of the archive, such as *"Elisabeth Järnefeltin arkisto"* and the name of the series or collection of letters, such as *"COLL.101.15."*.

To maintain provenance, we also preserve information on the actors as they are originally recorded, as well as the subset of the original dataset, such as the entire row in a CSV table. This approach ensures that any mistakes made during automatic processing can be detected and that the final user can view the original records. It enhances the trustworthiness of the harmonized dataset and allows for the easy reproduction of the process. In other words, our fully automatic process is transparent, and anyone can see the operations being performed. In case of the word files, we document the manual processing and once we have processed the data, we will publish a document listing all the changes that we have made. This part is the least transparent because it is challenging to show the original segments, but we hope that the manual will provide the user with a clear understanding of the cleaning that has been done.

We have an issue in dealing with families and institutions as actors in the data. For instance, some datasets only specify the recipient as "family Carlstedt", without providing details on the specific person within the family who received the communication. Letters could be written to the whole family or a group of family members and as we do not know if there were possible individual letter receivers in these cases, we have to retain the recipient as recorded. Likewise, when institutions are involved, we often lack information on the particular sender or recipient of a letter.

Structural issues

We have received certain datasets as data dumps from internal institutional database systems. However, the dumping process occasionally encounters errors, as evidenced by error reports found in some datasets, such as:

```
  No signature of method: com.zetcom.mp.service.provider.data.search.impl.
lucene.domainFacade.ControlledVocabularyNodeImpl.plus()  is  applicable
for argument types: (java.lang.String) values: [, ] Possible solutions:
is(java.lang.Object), split(groovy.lang.C
```

In addition, we have observed that some cells have been shifted left or right in a couple of CSV datasets. It is crucial to detect such errors and implement appropriate handling mechanisms to avoid obtaining anomalous results.

Names and persons without names

In some cases, we need to extract an actor's name from free text, as exemplified by phrases such as "Elisabeth Järnefelt's fonds", "Ensio Hiitonen's conversations" or "Aarno Durchman's letters to Sigrid Duchman" (originals in Finnish or Swedish). While obtaining the basic form of the name from the genitive case is relatively simple in English, it is much more complex in Finnish. In Finnish, words not only receive a suffix, but also undergo changes. For instance, here are several examples of genitive cases and basic forms in Finnish: "Järnefeltin – Järnefelt", "Suolahden – Suolahti", "Hiitosen – Hiitonen ".

To obtain the base form of names, we used the FinnPos lemmatizer [10]. The lemmatizer uses Finnish morphology Omorfi [11] but can also lemmatize unknown words, making it an excellent tool for handling names, particularly given the presence of foreign names in our data sets. Nonetheless, the lemmatizer may occasionally make mistakes, which is why we manually corrected the results and stored them in a dictionary for easy reuse.

Moreover, in all wide correspondences, there are also unknown senders, persons who have not been be identified by close-reading the letters or from other correspondence. Currently, we have identified approximately 30 different ways to label unknown individuals in datasets, with the term "Tuntematon" being the most frequent one. Having unidentified senders may be

Table 3
Example from the HS foundation CSV file. We are showing three rows and two columns, which contain details about senders and their sent letters. In the first row, an asterisk (*) indicates a telegram (sähke), while the letter S (S = saapuneet kirjeet) signifies that the listed letters are received ones. The recipient is the fonds' holder and that column is not displayed in this excerpt.

A cell in CSV file containing correspondence information	Other info
Nekton, Toivo H. (Broolyn, New York): S: 16.9.*, 20.9.* ja 13.10.1905	* sähke
Kirjeenvaihto Manda ja Juho Soinin kanssa: S: 25.7.1897.	S = saapuneet kirjeet
Kirjeenvaihto Edla Soldanin kanssa: S: 23.5.1870, 12.8., 4.9., 25.9., 21.10. ja 30.11.1872, 1.2., 7.3., 18.3., 1.4., 7.5., 28.5./16.6., 6.9., 9.10., 11.10. ja 1.11. 1873, Heikin päivänä, 1.3., 18.3., 2.4., 12.4., 2.5., 4.7., 23.7., 2.9., 13.9., 28.9., 7.10. ja 22.10.1874, 3.2., 4.4., 6.5., 24.7., 21.11.1875, 6.2., 26.3., 22.10.1876, 29.4., 1.7., 25.10., 30.12.1877, 22.9. ja 29.9. 1878, 1.4. ja 20.9.1879, 13.3.1880, 3.1. ja 5.5.1884. (5 nippua) Kirjeistä on myös valokopiot. (1 nippu)	S = saapuneet kirjeet

frustrating for the user of a single collection, but it becomes a problem when we bring together various fonds with numerous unknown senders. This means that each unidentified person must be given a unique and identifying "unknown identifier".

Parsing free text

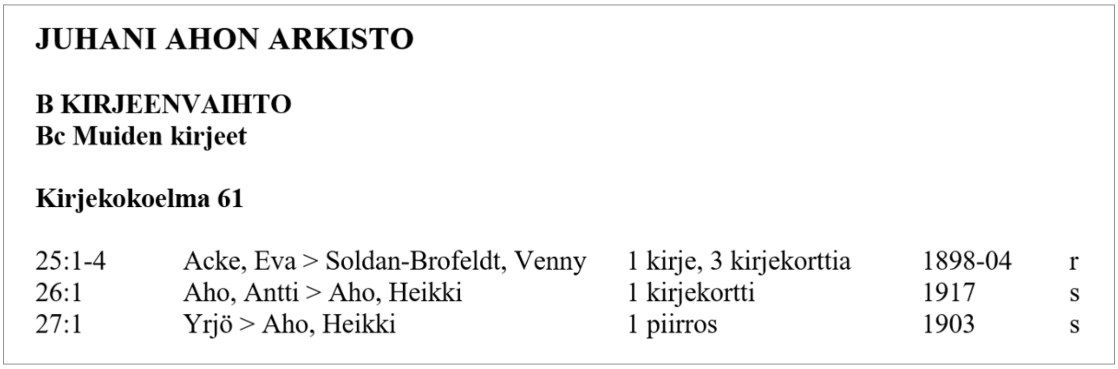
Parsing free text can be a challenge in certain datasets, even if they appear to be structured, such as in CSV format. While some examples, such as the National Gallery files, may contain sender and receiver information in a formatted text format (e.g., "Ingrid Carlstedt, sender; Mikko Carlstedt, recipient"), more complex cases may involve actors, places, and individual letter dates, with varying formats across cells. An example of such complexity is shown in Table 3. The first row indicates that the sender was "Nekton, Toivo H.", the place of sending was "Brooklyn, New York", and two telegrams (*sähke*) were sent in September 1905 along with one letter in October 1905. The second row involved two senders, "Manda and Juho Soini", who sent one letter. The third row shows that "Edla Soldan" sent multiple letters on different dates. The recipients of the letters are not displayed in this table, as that information was straightforward to process in a separate column.

The letter collection by the Finnish Art Society contained brief summaries of the letter content often mentioning related people, organizations or places. These summaries were written in a free text format, like for example "Walter Runeberg's scholarship to travel to Rome" or "Altarpiece for church in Kalajoki, Adolf von Becker". In processing NLP tools [12] based on FinBERT [13] were used to extract these references to named entities.

Processing information provided in this manner can be difficult, especially when dealing with large and varied datasets. To successfully parse the data, we must develop parsers to recognize different patterns, while also employing a lemmatizer to obtain the basic form of a name and using name processing algorithms to ensure consistency in the format of names.

Word files

Some prominent organizations like the National Library of Finland (NL) and the Swedish Literary Society in Finland (SLS) still maintain traditional word-format catalogues as the ”user interface” to their epistolary collections (in the case of NL, the letter metadata only exists in Word format). Such catalogues are perfectly suited to the needs of human users. They are sufficiently consistent to allow information seekers to quickly, on the basis of previous knowledge, to grasp their logic. However, they provide automatic, algorithmic reading with a set of specific challenges.



JUHANI AHON ARKISTO

B KIRJEENVAIHTO
Bc Muiden kirjeet

Kirjekokoelma 61

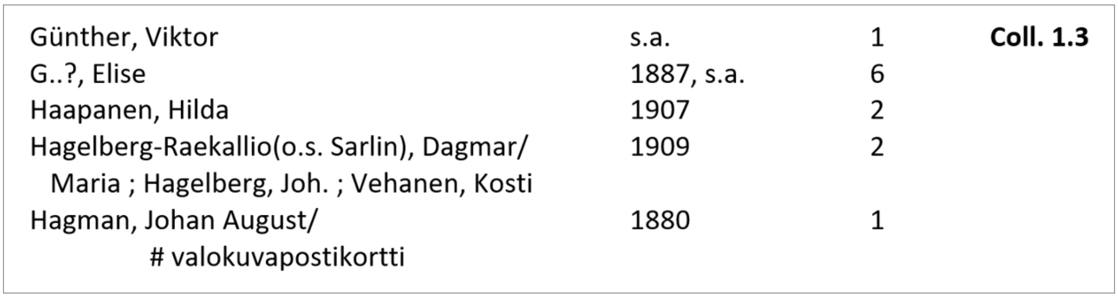
25:1-4	Acke, Eva > Soldan-Brofeldt, Venny	1 kirje, 3 kirjekorttia	1898-04	r
26:1	Aho, Antti > Aho, Heikki	1 kirjekortti	1917	s
27:1	Yrjö > Aho, Heikki	1 piirros	1903	s

Figure 4: Juhani Aho’s correspondence catalogue contains information about letter exchange (Kirjeenvaihto) between other people (Muiden kirjeet). Additionally, this particular catalogue includes information about the type of documents exchanged, such as letters (kirje), postcards (kirjekortti(a)), and drawings (piiros). The fonds of Aho is kept at the Finnish Literature Society, Helsinki.

The catalogues have been created by different archivists in a wide variety of organizations with their own cataloguing practices over the decades, resulting in inconsistent formatting. Typically these files begin with the name and brief biography of the records creator, the primary person whose archive it is, followed by information about the various documents in their archive. The catalogue section that is of particular interest to us is the Letter Exchange, which is usually categorized into subcategories such as Received Letters, Sent Letters, Letter Concepts (sometimes “Unsent Letters”), and Letter Exchange between “Other Individuals”. In other words, the archive (and the correspondence catalogue) of the prominent 19th-century author Juhani Aho, held in the archive of the Finnish Literature Society, contains letters exchanged by his wife Venny Soldan-Brofeldt and her acquitance Eva Acke, as shown in Figure 4. Except for the last subcategory, all the other subcategories usually contain four columns separated by tabs or spaces, with details about the name of an actor, time, quantity of letters, and the signum (the collection’s reference code). The section containing letter exchange between other individuals includes details about two actors, the sender and the recipient, along with the time, quantity, and the signum.

Parsing the files automatically poses several challenges due to various issues in the files. Firstly, the general document structure is diverse. Some catalogues contain archival information not only on one person but also on their spouse and other family members. Additionally, not all documents follow the same structure, and catalogues with abundant material often have a wide variety of formatting. The subsections of the letter exchange also do not have naming conventions, but require a deeper semantic language understanding due to their creativity.

Inconsistencies at the line level further compound the parsing challenge. For example, sometimes information about one letter correspondence spans multiple lines, either due to insufficient space or additional information such as comments or place of sending. In some cases, there is even information on some other correspondence between other people.



Günther, Viktor	s.a.	1	**Coll. 1.3**
G..?, Elise	1887, s.a.	6	
Haapanen, Hilda	1907	2	
Hagelberg-Raekallio(o.s. Sarlin), Dagmar/ Maria ; Hagelberg, Joh. ; Vehanen, Kosti	1909	2	
Hagman, Johan August/ # valokuvapostikortti	1880	1	

Figure 5: A sample of Ida Aalberg's collection of received letters from the National Library Word files, which has undergone manual editing.

Since the data is highly heterogeneous, a complex parser is required to process it. However, creating such a parser would be time-consuming and labor-intensive. Therefore, we have opted for a semi-automatic approach in which research assistants manually examine the documents to identify inconsistencies and harmonize the information. We have established a set of rules to manually transform the original dataset into a consistent format that can be parsed automatically. These rules include separating catalogues with information on multiple main archival persons into separate documents, harmonizing subsection titles, and using "/" to mark line breaks, "#" for general comments, and "##" for comments with additional information. We are also standardizing sections that contain information about correspondence between other actors by introducing a consistent format to replace the diverse formats found throughout the dataset. An edited file excerpt is presented in Figure 5, illustrating an example from Ida Aalberg's catalog. In this example, we have added "/" to mark line breaks, seperated authors with " ; " and added a "#" in front of a comment.

4. Data Model and Harmonization

In order to present the aggregated heterogeneous epistolary datasets in a coherent, unified format, work on developing a harmonizing data model for epistolary metadata collections is undergoing in the project. The CoCo data model builds on international standards like CIDOC CRM[11] [14], Dublin Core[12], and ICA Records in Contexts[13] to promote interoperability. The data model aims to support modeling of the relevant properties of letter metadata that we

[11] https://www.cidoc-crm.org

[12] https://www.dublincore.org/specifications/dublin-core/dcmi-terms/

[13] https://www.ica.org/en/records-in-contexts-ontology

have collected from the source datasets (see Table 2), to support efficient use of the data in e.g. SPARQL queries and the semantic portal developed during the project.

The current version of the data model includes the most central classes that are Letter, Production, Actor, Place, and Time-Span. Also, provenance (class MetadataRecord) and archival/collection level information (classes Series and Fonds) are included in the data model. During the transformation process, the intermediary RDF format is converted into RDF format corresponding to the CoCo data model. Instances of classes, such as Actor, Place, and Time-Span, are created based on the literal values of the intermediary format.

For representing actors (senders and recipients of letter) in different source datasets, we use an adaptation of the proxy concept from Open Archives Initiative Object Reuse and Exchange (OAI-ORE)[14]. In our case, a Proxy stands for a certain perspective on a person or group in the context of a specific source. In the harmonization process, proxies that are identified (by the future deduplication/disambiguation workflow) to represent the same person or group are connected using a shared instance of the class ProvidedActor. The class is an adaptation of the Europeana Data Model's[15] class ProvidedCHO (Provided Cultural Heritage Object). In Europeana, a ProvidedCHO "represents the Cultural Heritage Object that Europeana collects descriptions about".

The actor data is enriched by linking it to external databases like Wikidata and the Finnish AcademySampo [15] and BiographySampo [16]. These external sources provide detailed biographical information, e.g., times and places of birth and death, name variations, occupations, or genealogical relationships. Information present in the letter metadata like actor names and times of sending and receiving is used for matching entities between our data and the external databases, and further to reconcile the actors between data sources.

5. Discussion and Conclusion

Recent digital history discussions emphasise the moral or the ethical side central to the use of big data resources. [2, 17] Also, a distinction has been made between technical and ethical data work. [2]. The data transformation pipeline described above, with its different stages of manual, semi-manual and automated processing, can be understood as the technical side of such a project, sometimes referred to as "data cleaning". However, Katie Rawson and Trevor Muñoz have persuasively argued that we should avoid using this phrase. According to them, its (often slightly) offhand use implies that researchers regard the time-consuming data processing as having no tangible impact on the value of the research findings and therefore its detailed description has no relevance. Rawson and Muñoz use phrases such as "critically attuned data work" that enables us to see "the messiness of data not as a block to scalability, but as a vital feature of the world that our data represents and from which it emerges." [18]

The great variety of formatting and the combination of scalable and nonscalable elements in the Word files seemed at first to be a real obstacle for the data processing. However, we gradually realised that without the existence of this diversity we would not understand our data – also the parts we received for example in CSV format – as well as we currently do.

[14] https://www.openarchives.org/ore/

[15] https://pro.europeana.eu/page/edm-documentation

Moreover, it turned out that what we are doing is not merely ”cleaning up” other peoples’ mess as efficiently as possible. As we work through or with the data, we genuinely seek to understand the specificities of the data and consequently try to harmonise it in ways that do not lose the specificity of each correspondence. In this process, we create a new dataset (archive) with its own, regulated vocabulary.

The premises of the ethical data work ran parallel to Rawson’s and Muñoz’s discussions. According to the pivotal *The Network Turn* by Ahnert *et al.*, ethical data work aims at revealing gaps, biases and holes in data sets. The authors suggest that we should look at our data as a perspective, not as a bias. [2] Ethical data work demands a high-standard documentation of all the measures and the whole process of data work done in the project. We must be open towards the future users of our portal and data, and provide them with the information on the metadata that we have and how we have processed them. It also entails scrutinizing the data in the context of participating CH organisations and their collection histories and policies.

One interesting humanistic approach to discuss this (serendipitous and active) selection process is to frame it in terms of cultural heritage. When we do not reduce “cultural heritage” merely to its material manifestations but rather understand it as a dynamic process of making and becoming, we can conceptualize the analogue and digital cataloguing and describing processes as acts of active heritagization. Thus, the production of metadata constitutes an integral part of the discursive framework of values, meanings and relevance that define institutional heritage preservation. 19th-century letters have been potentially heritagizied when they have been included in the collections of CH organizations, and actively heritagizied when they have been provided with metadata. [19] One could perhaps argue that the work described in this paper adds yet another layer to this process. It re-heritagizes those epistolary fonds included in our dataset. They will become available and visible in a way that would not be possible in an analogue archive, and interwoven into the fabric of other – digitized, laboriously yet respectfully processed and connected – cultural heritage.

Acknowledgments

Our work was funded by the Academy of Finland as part of the project Constellations of Correspondence: Relational Study of Large and Small Networks of Epistolary Exchange in the Grand Duchy of Finland (CoCo) (decision numbers 339828, 340834, and 339918). CSC – IT Center for Science, Finland, provided computational resources for the work.

References

[1] J. Tuominen, M. Koho, I. Pikkanen, S. Drobac, J. Enqvist, E. Hyvönen, M. La Mela, P. Leskinen, H.-L. Paloposki, H. Rantala, Constellations of Correspondence: a linked data service and portal for studying large and small networks of epistolary exchange in the Grand Duchy of Finland, in: 6th Digital Humanities in Nordic and Baltic Countries Conference, short paper., 2022, pp. 415–423. URL: http://ceur-ws.org/Vol-3232/paper41.pdf.

[2] R. Ahnert, S. E. Ahnert, C. N. Coleman, S. B. Weingart, The Network Turn: Changing Perspectives in the Humanities, Cambridge University Press, 2020.

[3] M. Doerr, S. Gradmann, S. Hennicke, A. Isaac, C. Meghini, H. Van de Sompel, The europeana data model (edm), in: World Library and Information Congress: 76th IFLA general conference and assembly, volume 10, IFLA, 2010, p. 15.

[4] W. Ravenek, C. van den Heuvel, G. Gerritsen, The epistolarium: origins and techniques, CLARIN in the Low Countries (2017) 317–323. URL: https://doi.org/10.5334/bbi.26.

[5] A. Rockenberger, E. N. Wiger, M. R. Witting, H. Bøe, E. I. Thor, O. J. Wolden, M. Paasche, O. Søndenå, P. Conzett, Norwegian correspondences and linked open data, in: Proceedings of the Digital Humanities in the Nordic Countries 4th Conference, volume 2364 of *CEUR Workshop Proceedings*, 2019, pp. 365–375. URL: http://ceur-ws.org/Vol-2364/33_paper.pdf.

[6] C. van den Heuvel, Mapping knowledge exchange in Early Modern Europe: Intellectual and technological geographies and network representations, International Journal of Humanities and Arts Computing 9 (2015) 95–114. URL: http://doi.org/10.3366/ijhac.2015.0140.

[7] D. van Miert, What was the Republic of Letters? A brief introduction to a long history (1417–2008), Groniek 204/205 (2016) 269–287.

[8] H. Hotson, T. Wallnig (Eds.), Reassembling the Republic of Letters in the Digital Age: Standards, Systems, Scholarship, Göttingen University Press, 2019.

[9] O. Bruneau, N. Lasolle, J. Lieber, E. Nauer, S. Pavlova, L. Rollet, Applying and developing semantic web technologies for exploiting a corpus in history of science: The case study of the henri poincaré correspondence, Semantic Web 12 (2021) 359–378.

[10] M. Silfverberg, T. Ruokolainen, K. Lindén, M. Kurimo, FinnPos: an open-source morphological tagging and lemmatization toolkit for Finnish, Language Resources and Evaluation 50 (2016) 863–878.

[11] T. A. Pirinen, Omorfi—free and open source morphological lexical database for finnish, in: Proceedings of the 20th Nordic conference of computational linguistics (NODALIDA 2015), 2015, pp. 313–315.

[12] M. Tamper, A. Oksanen, J. Tuominen, A. Hietanen, E. Hyvönen, Automatic annotation service appi: Named entity linking in legal domain, in: A. Harth, V. Presutti, R. Troncy, M. Acosta, A. Polleres, J. D. Fernández, J. Xavier Parreira, O. Hartig, K. Hose, M. Cochez (Eds.), The Semantic Web: ESWC 2020 Satellite Events, volume 12124 of *Lecture Notes in Computer Science*, Springer-Verlag, 2020, pp. 208–213. URL: https://doi.org/10.1007/978-3-030-62327-2_36. doi:10.1007/978-3-030-62327-2_36.

[13] A. Virtanen, J. Kanerva, R. Ilo, J. Luoma, J. Luotolahti, T. Salakoski, F. Ginter, S. Pyysalo, Multilingual is not enough: BERT for finnish, CoRR abs/1912.07076 (2019). URL: http://arxiv.org/abs/1912.07076. arXiv:1912.07076.

[14] M. Doerr, The CIDOC CRM—an ontological approach to semantic interoperability of metadata, AI Magazine 24 (2003) 75–92.

[15] P. Leskinen, H. Rantala, E. Hyvönen, Analyzing the lives of finnish academic people 1640–1899 in nordic and baltic countries: Academysampo data service and portal, in: DHNB 2022 The 6th Digital Humanities in Nordic and Baltic Countries Conference, CEUR Workshop Proceedings, long papers, Vol. 3232, 2022, pp. 94–108. URL: http://ceur-ws.org/Vol-3232/paper07.pdf.

[16] M. Tamper, P. Leskinen, E. Hyvönen, R. Valjus, K. Keravuori, Analyzing biography collection historiographically as linked data: Case national biography of finland, Semantic

Web – Interoperability, Usability, Applicability 14 (2023) 385–419. URL: https://doi.org/10.3233/SW-222887.

[17] W. Kansteiner, Digital doping for historians: Can history, memory, and historical theory be rendered artificially intelligent?, History and Theory 61 (2022) 119–133.

[18] K. Rawson, T. Muñoz, Against cleaning, in: M. K. Gold, L. F. Klein (Eds.), Debates in the Digital Humanities 2019, University of Minnesota Press, 2019, pp. 279–292. URL: https://doi.org/10.5749/j.ctvg251hk.26.

[19] J. Enqvist, I. Pikkanen, Kirjeluettelot kulttuuriperintönä ja tutkimusaineistona: metadatan mahdollisuudet digitaalisen käänteen jälkeen (2023/2024). Accepted, forthcoming in 2023/2024.

Perspectives on sustainable dislocated digital research resources

Andrea Alessandro Gasparini[1], Tom Gheldof[2]

[1]*Department of Informatics, University of Oslo, Gaustadalléen 23 B, Oslo, N-0373, Norway*
[2]*Research Unit of Ancient History, KU Leuven, Leuven, Belgium*

Abstract

In difficult times, researchers often react promptly and adapt to a new situation, still focusing on the core values. In the upheavals of technology, researchers in Digital Humanities have also created innovative solutions. However, significant issues are associated with sustainable ingredients. First, software often emerges and disappears, leaving valuable data in the wild or causing costs in maintaining technological competences in the organization. Second, storing databases that have reached the end-of-life stage remains an unresolved issue. Third, knowledge and data can be closed inside a database and are often inaccessible to all. Finally, Digital Humanities must play a crucial role as a supporter of a sustainable world by rethinking its approaches. This paper uses ENCODE (an Erasmus+ funded project to bridge the gap between training and digital competences in the teaching/learning domain of ancient writing cultures) as a case study to examine some of the aforementioned issues.

Keywords

Digital Humanities, Sustainability, SDG, Systems Thinking

1. Introduction

Deconstructing and reconstructing knowledge are two of the major forces behind Digital Humanities (DH). Annotating old text to enrich forgotten languages, analyzing large bodies of images, or using artificial intelligence (AI) to predict missing parts of documents are a few of the various activities and technologies researchers are supporting today in the humanities in pursuing their tasks. During the last two decades, some radical changes have altered the technical infrastructure in which Digital Humanities may evolve. First, a fast and easy-to-connect network infrastructure allows anyone to contribute to digitization projects almost anywhere. Second, cheap computing resources are available at various sites, allowing the overproduction of projects and data. During this period, funding was easy to access, and a variety of small to very large projects started in academia and were occasionally supported by external software companies. As expected, this created a fragmented and complex landscape with several drawbacks [1]. For instance, it is very expensive to keep various projects alive, or

DHNB2023 | Sustainability: Environment - Community - Data. The 7th Digital Humanities in the Nordic and Baltic Countries Conference. Oslo – Stavanger – Bergen, Norway. March 8–10, 2023.

andreg@ifi.uio.no (A. A. Gasparini); tom.gheldof@kuleuven.be (T. Gheldof)
https://www.sustainabilitylab.uio.no/people/andrea/index.html (A. A. Gasparini);
https://www.kuleuven.be/wieiswie/en/person/00061049 (T. Gheldof)
0000-0002-1910-0859 (A. A. Gasparini); 0000-0001-7116-9338 (T. Gheldof)

DHNB Publications, DHNB2023 Conference Proceedings, https://journals.uio.no/dhnbpub/issue/view/875

data disappear [2]. Digital sustainability in DH addresses all of these aspects, from software, data models, and interfaces to long-term access and security issues [3]. For instance, a well-designed service interface with a rich and user-friendly visualization of knowledge will ensure and maximize the use of DH tools. In addition, when visualizing datasets, the "Interfaces become performative environments where scholars can play with the data and build their own interpretations." [4] Even though several activities in the context of digital sustainability do exist, what this paper emphasizes is an emerging, new, and critical role Digital Humanities need to expand and mature into: a supporter of a sustainable world by rethinking their approaches.

2. Digital Humanities for a sustainable world

The United Nations defined 17 Sustainable Development Goals (sdgs.un.org), covering issues, needs, and perspectives on poverty, gender, education, health, energy, consumption, production, peace, and climate change. In particular, problems concerning education, climate change, and energy consumption have implications for DH. Thus, the first question is how and why DH should consider the Sustainable Development Goals (SDG). What follows addresses the goals of developing long-term solutions for DH taking into consideration various SDGs and their relationship. Moreover, this paper will investigate DMP, FAIR, and Open Science [2] from a DH and SDG perspective.

2.1. Sustainability and software

DH projects use several technological infrastructures, including web software (e.g., front-end), sometimes a content management system, or a database platform (e.g., back-end). In addition, templates are necessary to register metadata in the system. The template is often a crucial part of the knowledge creation process, as it addresses the understanding and value of the data. When migrating to new platforms, reproducing the design process of building and testing the representation of data is quite difficult, and the lack of a "thick description" is often mentioned as a major risk for DH [5]. Preferably the interface has also been designed appropriately and tested with a User Experience (UX) approach.

All of the above can be developed in-house or by using off-the-shelf software. Nevertheless, all technological infrastructures have a life cycle [6], as software needs to be upgraded and, in the end, migrated to new ones. In addition, companies providing services may merge or disappear. Contextualizing the life cycle of technological infrastructures for DH in the academic world, the narrative is quite interesting. The anecdotal stories of professors with little funding, developing interesting services on computers under the desk in their offices, are often true and do not end well for the service.

Other issues, especially relevant for small projects, include security, as hacking into systems happens quite often and may destroy all the valuable data stored in the system. On the other hand, having goals of developing long-term solutions for DH projects already from the start, supports several SDGs. DH projects can foster innovation (SDG 9) when created in a professional environment. They support education for all (SDG 4) and reduce inequality (SDG 10) by supporting access to all the knowledge in the databases using, for instance, an Open Access approach when developing solutions for DH.

2.2. Sustainability and data

Data are another asset in DH, as they represent the work done during the research analysis, tuning, and cleaning of raw findings. Unfortunately, data are also stakeholders in DH, with several complicated aspects. Time is problematic for data stored in databases:first, occasionally one needs to migrate to new formats, and the long-term preservation of data needs to consider how various formats, such as XML (text) or PDF, resist migration needs [2]. Second, backups can easily become corrupted by large amounts of data [7]. From an SDG perspective, data are the tendo Achillis of DH projects. Both SDG 13 (climate action) and SDG 7 (arguing for affordable and clean energy) underpin the problematic aspects of producing, storing, and using data in DH activities. Energy consumption (SDG 7) is an emerging perspective on sustainability in digital humanities.

This is because storing and using data requires large amounts of power in data centers worldwide, but is invisible to users. For instance, images on websites, images, PDF, and other files are presented after a search, then may be downloaded locally, and finally stored in personal data houses (such as Google Drive, OneDrive, etc.), exponentially increasing the storage and transportation needs of the same data [8].

Another side effect of using energy and applicable to DH projects is called the 'rebound effect', where information of almost unlimited access to data storage increases the use, even the costs of consuming more energy affects the climate [9]. Finally, Digital Humanities projects should aim to reduce inequality (SDG 10). For instance, knowledge and data can be closed in a database and are often inaccessible to all. Therefore, researchers should be obliged to maximize the use of their data using the FAIR principles [10].

2.3. Sustainability and knowledge creation

The understanding of where and how new knowledge is created when the humanities interact with the digital world is understudied. When scholars started to develop DH services, they handed over the responsibilities to IT departments resulting in services "that have not adequately addressed the epistemological trajectories being designed into their technological infrastructures." [11]

Moreover, the making (practices) and thinking (theory) when developing DH services need to be addressed and designed accordingly [12]. Therefore, humanities scholars need to be aware of their role as developers and designers, as the knowledge inscribed in the interface and the service in general is relevant to how the content is used. From a sustainability perspective, the above addresses access to knowledge for all (SDG 4), where gender-based bias and other issues regarding inequalities based on race, ethnicity, and income (SDG 9) need to be considered.

In addition, the "outputs" of digital research projects in the humanities have specific values [13]. First, value is given by the understanding that users give to the output. Second, the preservation of the outputs over time includes, as mentioned, the design of the interface, especially how a search is developed. The latter is unfortunately volatile and undervalued.

3. ENCODE

n recent years, digitization of ancient written objects has become increasingly possible and sustainable because of the rapid development of e-infrastructures, digital infrastructures that provide services and tools in virtual (and collaborative) environments. In the framework of ENCODE, a three-year (September 1, 2020 - August 31, 2023) Erasmus+ Strategic partnership for higher education, partners from 6 European universities (Alma Mater Studiorum Università di Bologna, Julius Maximilian Universität Würzburg, KU Leuven, Università degli studi di Parma, Universität Hamburg, and Universitetet i Oslo) joined forces. They have - and are still -undertaking several steps towards bridging the existing gap in the teaching/learning domain of ancient writing cultures between the peculiar humanistic training and the now essential digital competences required for study, research, and employment.

3.1. ENCODE and Sustainability

The project's sustainability is guaranteed by compliance with several best practices. In the conception phase, the project was described in detail, using the template of the Erasmus+ funding application. In this application, 6 Intellectual Outputs were outlined, including the organization of multiplier events, conferences, and training events, hosted by each of the partner institutions. Although sustainability - a word that occurs 19 times in the proposal - was not directly linked to the SDGs, several outputs tried to overcome some of the goals and challenges presented in the document from the United Nations. SDG 4 (Quality Education) and SDG 9 (Industry, Innovation and Infrastructure) are implicitly included in some of the project's outputs, such as the framework of digital competences for students and teachers dealing with written cultural heritage or the design of innovative and customizable teaching modules for an ENCODE database, guidelines, and an online course. Moreover, the teaching methodology focuses on mutual learning among trainees and trainers, allowing replicability of activities by producing models of training sessions with different ENCODE modules and making self-training materials accessible to academics and other researchers [14].

Additionally, the project aims to adhere to the FAIR principles [10]. The ENCODE outputs such as the database and the online course should be findable via different platforms. Both outputs ensure the accessibility of the teaching modules and training materials by offering them freely available in open access and with an appropriate license. Such a license (CC BY or CC0) also guarantees the uptake and recycling of the modules, especially the training materials, making it possible for anyone to modify them according to their own needs [15]. Finally, by complying with the main standards in the field of ancient written materials (EpiDoc, Linked Open Data, etc.), it is possible to use all the ENCODE outputs in a collaborative manner and link them to most existing projects dealing with the study of ancient texts.

3.2. The ENCODE online course on #dariahTeach

One of the main outputs consists of the creation of a MOOC or online course, a collaborative platform on which introductions for teaching staff, researchers and other users will be brought together with the training materials that are produced for the different ENCODE training events. A key concern of the partners was the long-term sustainability of such a digital platform. Whereas in a first phase, the partners explored the possibilities of setting up their own platform (e.g., Blackboard, Coursera, etc.) or choose an existing commercial initiative (such as Udacity, edX or Google Classroom), regarding the long-term preservation of the materials, it was decided to team up with a partner from another European consortium, DARIAH.

The #dariahTeach platform, developed using the open-source software Moodle, provides academics and other partners from the GLAM sector with the opportunity to create a free course. In contrast to an actual MOOC, these online courses are more modular and do not have a fixed start date. Modular course design is an approach to course development in which course materials are organized into smaller, self-contained units that can be rearranged, updated, or replaced, as needed.

This approach can facilitate updates and revisions of the content over time, allowing for the incorporation of new research findings and ensuring that the course remains current and relevant. This can help ensure that the ENCODE course remains up-to-date, and provides learners with the most current and accurate information. Partnerships and collaborations with other organizations, such as DARIAH, but also domain-specific partners, such as the digital classicists community, can help promote the sustainability of such online courses by ensuring their long-term impact and availability. By exploring such partnerships and collaborations with other organizations, course creators can identify new opportunities for sustainability and impact, ensuring that the course remains relevant and valuable over the long term [16].

3.3. Challenges and issues

During the project's lifecycle, the ENCODE partners faced several challenges and issues regarding the sustainability of the different outputs. When, at the beginning, opting to host dislocated digital resources, the ENCODE team considered many different solutions. Regarding a sustainable approach, especially in trying to comply with the FAIR principles and best practices from Open Science, we decided to:

- Reduce the risk of obsolescence: the online course is hosted with an open-source technology such as Moodle and maintained by the #dariahTeach project. Partners develop course materials, in close collaboration with the research community. The ENCODE guidelines on different digital competences in the field of ancient writing cultures will be published in an open GitHub repository and in collaboration with the digital classicists community.
- Improve interoperability: ENCODE uses open-source technologies (e.g., eXist-db for the database platform) and standards compatible with other existing platforms and projects. By using and stimulating Linked Open Data, it is possible to re-use and modify the ENCODE training materials according to each user's needs.

However, some challenges have not been sufficiently addressed or have occurred during the project's lifecycle and are still being investigated:

- Reduce environmental impact: in addition to the use of open-source technologies and standards, ENCODE was and still is dependent on the support of the technological services from each of the partner's institutions. In many of the partner universities, reducing the environmental impact of digital research infrastructures is on the long-term agenda, but not a top priority for the short term.

- Hosting and the long-term preservation of dislocated digital resources: one of the partners (Hamburg University) engaged from the beginning of the project to self-host the ENCODE database with the teaching modules, but due to a career change from the local expert, Hamburg University even decided to completely opt-out of the Erasmus+ project, leaving us with the challenge to find a new, long-term hosting plan for the ENCODE database. For the other outputs, it was decided to collaborate with and trust existing platforms, such as GitHub as a repository for the guidelines, institutional repositories for the project's publications, and #dariahTeach for the online course.

Most of these challenges and issues had already been considered at the start of the project and described in the project proposal, although some decisions had to be reconsidered when facing the issues described above. This proves again how, in the lifecycle of (academic) research projects) also sustainability is an aspect that requires sufficient consideration beforehand, but also needs to be re-investigated along the way. The challenges of creating dislocated digital resources in the ENCODE project were (and are still) addressed during the monthly transnational partner meetings and in close collaboration with external partners.

4. Mapping Digital Humanities and sustainability

Understanding sustainability in Digital Humanities is complex. Researchers, students, IT and university staff, and the public are all stakeholders, as well as the IT infrastructure, including the software, the project with data, and how it is organized. One way to approach all complex connections and dependencies is to adopt a holistic approach rather than focusing on each separate part. Systems Thinking is one way to create an overview as it addresses systems as "interrelated, or interdependent parts that form a complex and unified whole that has a specific purpose" [17]. In addition, Systems Thinking emphasizes an understanding of systems without the aim of solving inherent problems [18]. Therefore, the integrative properties of systems thinking fit to approach complex problems, as is often the case in DH projects, with multiple stakeholders and divergent perspectives, without complicating the process. In addition, the richness of stakeholders in DH projects provides an arena where reorganization and conservation of tensions between the activities is the rule. For example, the ENCODE project encountered a situation where the IT staff in charge of the database service changed workplaces, creating an unstable situation with several downtime periods. In Systems Thinking, these tensions are called self-organization [19]. Self-organization is when a system acquires its own "identity", where "coherent patterns of relationships are internally structured and develop over time." [19]

4.1. Iterative Inquiry

As a method for mapping out and analyzing the structure and processes of complex systems, the authors opted to use the design-based method "iterative inquiry" [20, 18, 21]. This method explains the different levels in a social system and how all the components are connected and interrelated at the micro - meso - macro level. For instance, for the ENCODE project, the micro level includes the local staff of the project, while the meso level consists of the university and the national epigraphic (or other domain-specific) community. Finally, at the macro level are the international communities and the public at large. However, for instance, at the meso level, the ENCODE project also has a role of sharing and disseminating ancient writing cultures. To map out the entire system, the approach uses several nested undersystems, such as Function, Structure, Process, and Structure/Purpose. More specific:

- Function = the action activating the sequences in the system
- Structure = stakeholders and their relationship
- Process = activities
- Structure/Purpose = unique environment

4.2. Using "Iterative Inquiry" to analyse the ENCODE project

Using the web-based visual platform Miro to visualize the output of the "iterative Inquiry" process, the entire system design of the ENCODE project, it was possible to discover various invisible aspects. Figure 1 shows how the linearity of the ENCODE project, typical for a DH project, has changed. The circular understanding of the different levels of the project and how they are connected and interrelated visualizes the dependencies. What the iterative inquiry shows is an awareness of each component at the different levels. Moreover, when observing how the project functions between the micro and the macro level, part of the less relevant technical aspects disappear.

Another aspect emerging from the iterative inquiry emphasizes the difficulty of materializing the SDGs in DH projects. One experience of the ENCODE project is the rationale for addressing these perspectives from the very beginning. As mentioned earlier, the ENCODE project considers SDG 4 (digital illiteracy) and SDG 13 (climate action) regarding open access, as they were well addressed in the original proposal. In addition, several UX tests are planned to guarantee SDG 10 (inequality) and SDG 4 (digital illiteracy).

On the other hand, the energy crisis, SDG 7, was not easy to react to, as the project does not have a direct connection to service providers, for instance, where and how videos are stored and accessed. This section outlines an iterative inquiry that has not yet been used in the ENCODE project. However, it has the potential to guide the project in assessing the SDG related to energy use, as it involves actions on multiple levels.

5. Discussion

Digital sustainability has become a crucial element in the development of new Digital Humanities projects. The FAIR and Open Access movement does help address some of the issues of long term access to all users, and directly underpins the necessity to react to inequality (SDG 10) and digital illiteracy (SDG 4). Both digital sustainability and the role DH plays in supporting the goal of a sustainable world need to be seen and used simultaneously in DH projects. A call to rethink DH approaches used when starting a project was chosen for the ENCODE project, resulting in a dislocated placement of resources. The use of the iterative inquiry tool to create an overview and understanding of the ENCODE project (see Figure 1) is valuable. The Systems

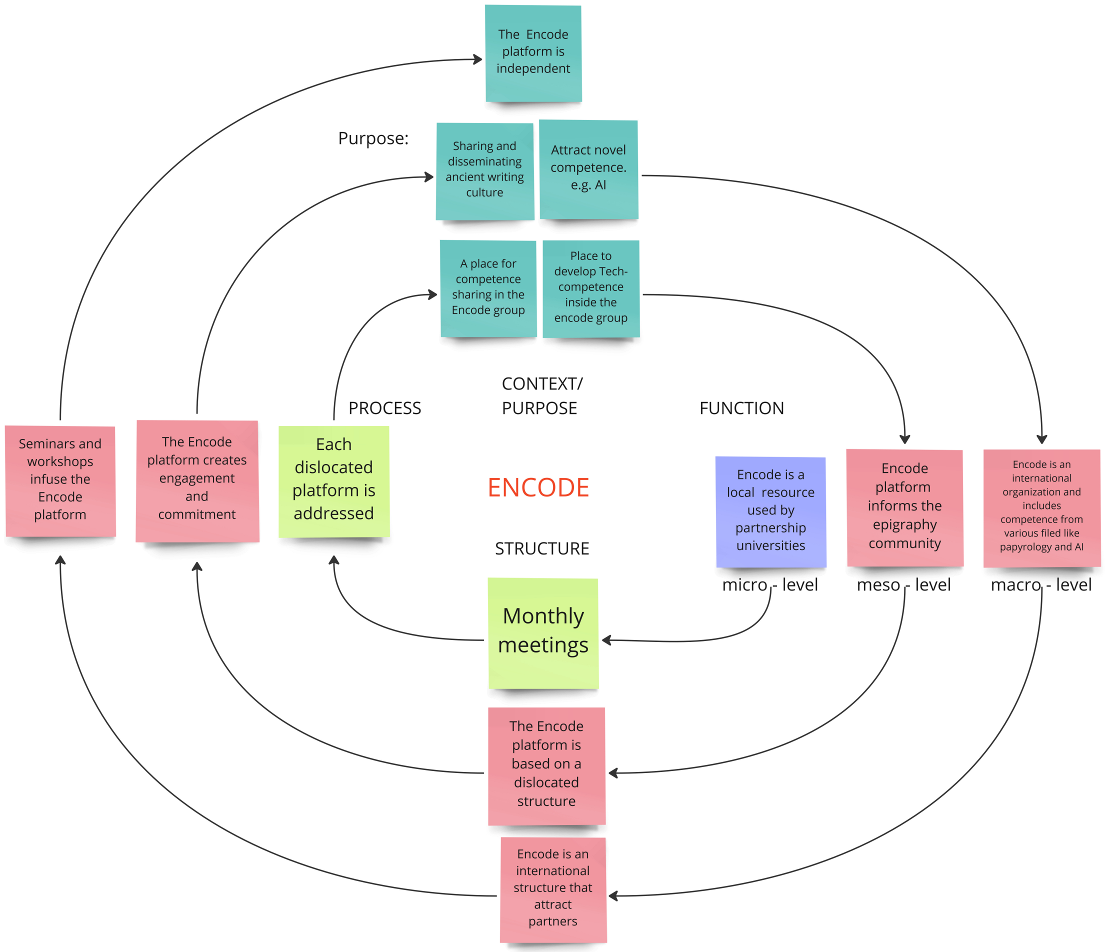

The Encode platform is independent
Purpose:
Sharing and disseminating ancient writing culture
Attract novel competence. e.g. AI
A place for competence sharing in the Encode group
Place to develop Tech-competence inside the encode group
CONTEXT/ PURPOSE
PROCESS
FUNCTION
Seminars and workshops infuse the Encode platform
The Encode platform creates engagement and commitment
Each dislocated platform is addressed
ENCODE
STRUCTURE
Encode is a local resource used by partnership universities
Encode platform informs the epigraphy community
Encode is an international organization and includes competence from various filed like papyrology and AI
micro - level
meso - level
macro - level
Monthly meetings
The Encode platform is based on a dislocated structure
Encode is an international structure that attract partners

Figure 1: Systems Thinking - contextual inquiry representation of the ENCODE project

Thinking approach and the tool can be used to examine tensions, such as the goal of the project versus sustainability or long-term preservation of data versus presenting earlier results of a project. These new scopes have changed the linearity of the ENCODE project toward a circular understanding of the different levels.

The emerging dependencies at the higher level (macro) of the iterative inquiry of the technical aspects of the project have mostly disappeared. This is an aspect that the ENCODE project is currently working on. In other DH projects, the focus is often on specific phases, such as results and findings and writing articles. While other aspects are often left behind, one example is the lifecycle of DH projects. This highlights the complexity of DH projects, from an advanced network of online services and infrastructures to proprietary project metadata. Therefore, the ENCODE project has opted for dislocated platforms such as GitHub for the Guidelines, and #dariahTeach for the MOOC online course.

Finally, energy consumption and other material usage when storing data or running servers for applications are affected by DH and need to be scrutinized with the SDGs in mind. For instance, SDG 7, affordable and clean energy, is based on a forecast that ICT will demand 21% of all energy produced in the world by 2030 (sdgs.un.org). When planning DH projects, the use of energy in data centers, which includes cloud computing, big data analysis, and Artificial Intelligence, also needs to consider the effects on resource extraction, manufacturing, transport, use, and end-of-life of ICTs. As mentioned before, how images, pdf, and other documents are stored and presented online may have an impact on energy consumption and indirectly on CO2 emissions (SDG13: Climate Action).

6. Guidelines

The following guidelines, supporting sustainable digital research resources, are the results of the ENCODE project. However, the last three emerged during the final phase of the project.

- Reuse of data (create a Data Management Plan)
- Public and easy access for researchers (Open Science)
- Support Linked Open Data
- Use the User experience (UX) approach so your data is user-friendly, and new knowledge is created and stored in the interface
- Use UD – Universal Design (e.g., for the visually impaired) access to all
- Decrease power usage by using links to images. Don’t copy them!
- Be aware that we must accept the scientific value of creating an open-access database!

7. Conclusion

This paper presents how the European project ENCODE, under the ERASMUS+ umbrella, has focused on sustainable research infrastructure. Approaching a project with a sustainable, dislocated digital resources mindset is possible. From the beginning, the sustainable aspect and the 17 Sustainable Development Goals (defined by the United Nations) have been addressed when planning the activities and as part of workshops and seminars that are part of the project’s outputs. These outcomes include guidelines to support sustainable digital research resources.

Acknowledgments

The work presented here is part of the Erasmus+ ENCODE project. The authors wish to thank all the project partners in ENCODE and are also grateful to the sustainabilitylab.uio.no for their support.

References

[1] K. Stapelfeldt, S. Khera, N. Ledchumykanthan, L. Gomez, E. Liu, S. Dhaliwal, Strategies for Preserving Digital Scholarship / Humanities Projects, The Code4Lib Journal (2022). URL: https://journal.code4lib.org/articles/16370.

[2] C. Barats, V. Schafer, A. Fickers, Fading Away... The challenge of sustainability in digital studies, Digital Humanities Quarterly 014 (2020). URL: http://www.digitalhumanities.org/dhq/vol/14/3/000484/000484.html.

[3] J. Tucker, Facing the challenge of digital sustainability as humanities researchers, Journal of the British Academy 10 (2022) 93–120. URL: https://www.thebritishacademy.ac.uk/publishing/journal-british-academy/10/facing-the-challenge-of-digital-sustainability-as-humanities-researchers/. doi:https://doi.org/10.5871/jba/010.093.

[4] S. Ramsay, On Building, in: Defining Digital Humanities, Routledge, 2013, p. 4. doi:doi:https://doi.org/10.4324/9781315576251, num Pages: 4.

[5] E. Tóth-Czifra, 10. The Risk of Losing the Thick Description: Data Management Challenges Faced by the Arts and Humanities in the Evolving FAIR Data Ecosystem, in: J. Edmond (Ed.), Digital Technology and the Practices of Humanities Research, Open Book Publishers, 2020, pp. 235–266. doi:10.11647/obp.0192.10.

[6] J. Smithies, C. Westling, A.-M. Sichani, P. Mellen, A. Ciula, Managing 100 Digital Humanities Projects: Digital Scholarship & Archiving in King's Digital Lab, Digital Humanities Quarterly 013 (2019).

[7] C. Taylor, Silent Data Corruption, the Backup Killer, Enterprise Storage Forum, 2016. https://www.enterprisestorageforum.com/management/silent-data-corruption-the-backup-killer/.

[8] J. Morley, K. Widdicks, M. Hazas, Digitalisation, energy and data demand: The impact of Internet traffic on overall and peak electricity consumption, Energy Research & Social Science 38 (2018) 128–137. doi:10.1016/j.erss.2018.01.018.

[9] V. C. Coroama, F. Mattern, Digital Rebound - Why Digitalization Will not Redeem us our Environmental Sins, in: Proceedings of the 6th International Conference on ICT for Sustainability(ICT4S 2019), volume 2382, RWTH, 2019, p. 31. URL: https://www.research-collection.ethz.ch/handle/20.500.11850/387584, accepted: 2020-01-14T16:34:23Z ISSN: 1613-0073.

[10] S. Europe, FAIR Principles, 2021. URL: https://www.go-fair.org/fair-principles/.

[11] A. Burdick, H. Willis, Digital learning, digital scholarship and design thinking, Design Studies 32 (2011) 546–556. doi:10.1016/j.destud.2011.07.005.

[12] G. Caviglia, P. Ciuccarelli, N. Coleman, Communication Design and the Digital Humanities Visualizations and Interfaces for Humanities Research, in: Proceedings of the 4th International Forum of Design as a Process, Brasil, 2012, p. 9.

[13] J. Edmond, F. Morselli, Sustainability of digital humanities projects as a publication and documentation challenge, Journal of Documentation 76 (2020) 1019–1031. URL: https://www.proquest.com/docview/2431709642/abstract/8EDA01110E3145F1PQ/1. doi:10.1108/JD-12-2019-0232, num Pages: 13 Place: Bradford, United Kingdom Publisher: Emerald Group Publishing Limited.

[14] Erasmus+, ENCODE (Bridging the <gap> in Ancient Writing Cultures: ENhance COmpetences in the Digital Era, Erasmus+, 2023. https://erasmus-plus.ec.europa.eu/projects/search/details/2020-1-IT02-KA203-079585.

[15] C. Commons, Creative Commons, About CC Licenses, Creative Commons, 2023. https://creativecommons.org/about/cclicenses/.

[16] C. Salvaterra, A. Bencivenni, M. Fogagnolo, T. Gheldof, I. Vagionakis, Encode4openu and the Preparation and Delivery of an International Collaborative MOOC: A Preliminary Analysis of its Pedagogical and Technical Implementation, Education Sciences 13 (2023). doi:10.3390/educsci13010043.

[17] H. Kim, Daniel, Introduction to Systems Thinking, Pegasus Publishing, Boulder, Colo, 1999. URL: https://thesystemsthinker.com/wp-content/uploads/2016/03/Introduction-to-Systems-Thinking-IMS013Epk.pdf.

[18] P. H. Jones, Systemic Design Principles for Complex Social Systems, in: G. S. Metcalf (Ed.), Social Systems and Design, Translational Systems Sciences, Springer Japan, Tokyo, 2014, pp. 91–128. URL: https://doi.org/10.1007/978-4-431-54478-4_4. doi:10.1007/978-4-431-54478-4_4.

[19] J. J. Kay, An introduction to systems thinking., in: D. Waltner-Toews, N.-M. E. Lister, J. J. Kay (Eds.), The Ecosystem Approach: Complexity, Uncertainty, and Managing for Sustainability, Columbia University Press, New York, 2008, pp. 3–34.

[20] P. Jones, K. v. Ael, Design journeys through complex systems: practice tools for systemic design, BIS Publishers, Amsterdam, 2022.

[21] P. Jones, J. Bowes, Rendering systems visible for design: Synthesis maps as constructivist design narratives, She Ji: The Journal of Design, Economics, and Innovation 3 (2017) 229–248. doi:10.1016/j.sheji.2017.12.001.

Nature and Culture in the Age of Environmental Crisis: Digital Analysis of a Global Debate in *The UNESCO Courier*, 1948-2020

Benjamin Martin [1], Fredrik Mohammadi Norén [2]

[1] *Uppsala University, Box 629 751 26 UPPSALA*
[2] *Malmö University, Box 50500 202 50 Malmö*

Abstract

This study uses digital text analysis, focusing on LDA topic modeling, to conduct a historical investigation of the relationship between the concepts of nature and culture found in the pages of the official magazine of the United Nations Educational, Scientific and Cultural Organization, *The UNESCO Courier*, between 1948 and 2020. The relationship between the concepts of nature and culture has historically been at the core of concerns about the environment and sustainability; *Courier* offers a means of charting a global conversation on these concepts. After presenting the corpus and our methods, the paper documents three approaches to LDA topic modeling that we have tested, through which we seek to make topic modeling useful for the field of conceptual history. Our empirical findings suggest that the concepts of nature and culture have come to be increasingly close over the course of the last six decades, while the stakes of the very distinction between the concepts have changed radically. Our methodological tests support the argument that topic modeling can be a valuable tool for conceptual history, albeit one that must be handled with care.

Keywords

Conceptual history, Global history, digital text analysis, topic modeling, UNESCO

1. Introduction

UNESCO is the United Nations organization for education, science and culture — not, in the first instance, for the environment or nature. But the organization has a long history of concerning itself with issues related to the natural environment. UNESCO's monthly magazine *Courier* reveals a striking level of interest in such issues. Since the magazine's foundation in 1948, *Courier* has featured articles that documented diverse ways that the world's peoples live in their natural environments, explained how particular cultures shaped landscapes, presented breakthroughs in scientific knowledge about nature, celebrated efforts to preserve humanity's "natural heritage", and, more recently, discussed the role of human activity in changing the Earth's climate [1]. At the heart of each of these topics was a set of fundamental questions about the relationship between nature and culture.

In the history of thinking about sustainability, the environment and related issues, the shifting and competing understandings of the relationship between nature and culture constitute a fundamental element. Indeed, the nature-culture dichotomy is a classic theme in the history of concepts. After all, as historians have shown, the concepts of culture and nature have always existed in relation to one another: defining a sphere of human autonomy and self-fashioning (culture) requires an opposing category encompassing that which humans did not create and which they struggle to control (nature), and vice versa [2]. The discussions of nature in *Courier*, a journal devoted to the themes of education, science and culture, offers an interesting source for exploring the history of that dichotomy in a dramatic phase in world history, during which the natural environment came to be a topic of international concern [3].

The 7th Digital Humanities in the Nordic and Baltic Countries Conference (DHNB 2022), Bergen, Norway, March 8–10, 2022
EMAIL: benjamin.martin@idehist.uu.se (B. Martin); fredrik.noren@mau.se (F. Mohammadi Norén)
ORCID: 0000-0003-3180-2973 (B. Martin); 0000-0001-8820-1082 (F. Mohammadi Norén)

DHNB Publications, DHNB2023 Conference Proceedings, https://journals.uio.no/dhnbpub/issue/view/875

The presence of nature-related themes in *Courier* is of particular interest, moreover, because of the publication's global character. Founded to "promote UNESCO's ideals, maintain a platform for the dialogue between cultures and provide a forum for international debate," *Courier* had uniquely global aspirations and reach. At its high point in the 1970s and 1980s it featured articles from prominent intellectuals across the globe published in 35 languages with an overall distribution of over 1.5 million copies, and was available on both sides of the Iron Curtain [4]. This magazine was recently digitized and made available by UNESCO (en.unesco.org/courier/archives). Working with developers at Humlab (Umeå University), we are curating this archive into a machine-readable corpus suitable for digital text analysis.

In this paper, we use *Courier* to follow a global conversation on the relationship between nature and culture. Given the magazine's focus on cultural matters, one means of charting that relationship is by focusing on the uses of the concept of culture therein. We do this by deploying tools of digital text analysis, in particular LDA topic modeling, in combination with close reading, to locate contexts in which the concept appeared in the magazine, measure how connections between these contexts changed over time, and identify novel ways to chart changes in the way the nature-culture relationship was articulated.

On the basis of these three approaches to using topic modeling we make a set of empirical findings and a methodological argument. In empirical terms, our preliminary findings allow us to identify thematic contexts in which nature was discussed in *Courier*, as well as to rank those contexts in terms of their relative strength in the corpus. They also reveal a fundamental debate in the publication over how to think about the concept of "nature" – a debate that advanced through arguments about what the relationship was (or should be) between nature and culture. Regarding methodology, we argue that topic modeling can serve as a useful tool for conceptual history, provided one is careful about what each computer-generated "topic" is, and is not.

2. Corpus and Method

The source for this study is the text printed in *Courier* from the magazine's foundation in 1948 until 2020. The print run was quite consistent for most of that time: the magazine was published monthly from 1948 until 2002, at which point *Courier* decreased the number of issues per year. After 2002, the publication rate changed often, and ceased entirely from 2013 to 2017, at which point it was relaunched at a rate of four issues per year. We downloaded the PDFs and used the open-source Tesseract OCR Engine to make the text machine readable. While not perfect, the OCR quality is good. Our own test of samples from the whole corpus generated an error rate of only 0.7 %. The total *Courier* corpus, from 1948 to 2020, consists of some 13 million tokens.

The most significant feature of this corpus is that, in contrast to large, general-language corpora (such as Google Books or Eighteenth-Century Books Online), ours is a focused, historically specific, and carefully curated text corpus. It is more specific in the themes it addresses than many other digitized publications, like daily newspapers, and it is linked in a specific way to a particular social agent: not just a publisher, but an international organization. This allows us to pose and hopefully answer questions of a sharper kind than those one can study through a broad, generic corpus, while also allowing us to link those questions to a social reality beyond the words we find in *Courier* – namely, the history of UNESCO as an institution. Moreover, and most importantly, insofar as *Courier* sought to give voice to representatives of UNESCO's member states, this is a corpus that gives us access to something approaching a global conversation – something that very few other bodies of text can match.

Our method in this investigation focuses on LDA topic modeling. This is a probabilistic method suitable for structuring a large and diverse text collection. Based on word distributions in documents, the model assigns each word a probability value and structures them into topics, and packages the underlying documents into blends of topics. Topic modeling builds on the principle that a word can be part of several topics, and that all topics are distributed in every document, but with different degrees of probability, and sometimes with very low values. There are different topic modeling algorithms, and this article uses the popular Latent Dirichlet Allocation, as implemented in Mallet [5]. The researcher decides on the number of topics to include, but after that, the model works unsupervised – inductively, one could argue – to identify "topics": top lists of words that have a higher probability of occurring close to one another in the different documents.

The researcher can then interpret and label these "topics" to capture the theme that each collection of words seems to reflect – often by moving back and forth between the statistical results (the lists of words) and passages from the documents themselves, in order to identify the sense of these topics as clearly as possible. Usually, several topics are more or less easily interpretable, while others are quite opaque. In the best case, allowing LDA to identify these topics offers "a method for analyzing texts...that is substantively quicker, more efficient and more objective than traditional methods of content analysis in the social and cultural sciences" [6]. At the same time, care is called for in handling these "topics", particularly for historical research. Each topic identifies real features of the text, identifying collections of words that may (in the best cases) correspond to what linguists (and some conceptual historians) call a "semantic cluster". But computer-generated topics cannot necessarily be understood as a "theme" or "discourse" in the sense in which these terms are use in the humanities and social sciences, much less as a "concept". If we are clear about these limits, however, the ability of the method to offer a fresh take on a large corpus has advantages for the history of concepts that we aim to explore here.

Before producing the topic models of *Courier*, some corpus preparations were made. We performed a part-of-speech tagging of the corpus and excluded numerals, delimiters, and all tokens that occur less than five times in total. A single page was chosen as the "document level" for creating the topic model. Our text curating will soon enable us to analyze at the level of articles, instead. For now, we used UNESCO's own index of *Courier* articles to automatically identify and extract the content of each issue from the first page on which an indexed article appears to the last. This excludes non-article content such as publication information (the publisher's imprint or "masthead"), tables of contents, editorials, and letters to the editor. *Courier* is a magazine free from traditional advertisements, and thus mainly consists of editorial content, including articles, photos with captions, and sometimes tables and graphs.

Having computed several different models, we selected a model of 200 topics (calculated on the basis of the corpus as divided up into single pages). We interpreted these topics manually and assigned each one a thematic label. A Jupyter notebook environment was set up to explore the topics through different tools, including topic word distribution, topic over time, and topic networks.

To use topic modeling for this historical investigation, we divide *Courier*'s print-run into sub-periods, so as to be able to observe change over time. Inspired by the periodization used in Cholé Maurel's history of UNESCO, which identifies periods corresponding to the tenures of the Directors General, we divide up the corpus into the following three blocs [7]:

Phase 1: 1945–1961: From Huxley to Veronese (16 years)
Phase 2: 1962–1986: Maheu and M'bow (24 years)
Phase 3: 1987–2019: From Zaragoza to Azoulay (28 years)

These periods correspond moreover to broad changes in UNESCO's history. It was in the early 1960s that the organization's membership expanded with the accession of many postcolonial states; the mid-1980s is widely seen as having marked a political-ideological transition, occasioned not least by the withdrawal from UNESCO at that time of the United States, Great Britain, and Singapore [8].

The methodological core of this article is an effort to use LDA topic modeling to explore the history of certain concepts. We seek thereby to contribute in particular to the recent research in transnational and global conceptual history, as well as to the scholarship in what can be called digital conceptual history [9]. Work in this emerging field has explored a variety of methods, including word-trend and collocate analysis, word embeddings (or vector-space models), as well as network analysis of co-occurrence data [10]. What can topic modeling offer to these efforts? It is by now relatively uncontroversial to claim that topic modeling, by providing "a way for researchers to obtain reasonable automated content coding of large text corpora," can offer a stimulating means of exploring a large body of text, insofar as the algorithm often suggests perspectives and connections that would otherwise be missed [11]. But this is still essentially a means of determining *what* is in the corpus. Here we ask whether topic modeling can help us understand not just *what* was discussed in *Courier*, but *how* it was discussed—and how that discussion changed over time.

In what follows we use topic modeling in three ways:

1. We apply word searches to our topic model as a means of identifying thematic contexts in which the concepts under examination appeared, and in order to explore the relative strength of those contexts in the corpus.
2. We apply network analysis to our topic model in order to identify relationships among contexts in which the concepts appeared.
3. We use the topic model to identify groups of pages in the magazine in which a chosen topic (which contains our target concept words) was particularly strong in a selected time period, for manual reading and analysis.

3. What did *Courier* talk about when it talked about nature?

To examine discussions of nature in *Courier* we need first to locate these. To do that, we identified all the topics (of the 200 generated by the algorithm) in which the words ‘nature’ or ‘natural’ appear among the top 50 words. One could of course simply search for all appearances of the words ‘nature’ and ‘natural’ in the corpus. The virtue of instead following these words’ appearance in a topic model is that doing so leads us immediately to a set of thematic foci, which one would otherwise need to build up from scratch. The themes suggested by the topic modeling output are, of course, not themes that we designed or selected in advance. Topic modeling serves us, then, as a data-driven means of identifying the contexts within which nature was discussed in *Courier* in a manner that is able to generate discovery and surprise from the very beginning.

Proceeding in this manner identifies twenty unique topics, about which we can make several observations.[2] Each “topic” consists, of course, simply of a ranked list of words. Happily, many of these topics are readily interpretable and seem easy to name. For example, topic 106 (“park”, “species”, “conservation”, “biosphere”, “national”, “reserve”, “natural”, “reserves”, “nature”, “world”), clearly has to do with national parks and other types of nature reserves, and is thus labeled “national parks”. Topic 50 (“earthquake”, “disaster”, “earthquakes”, “damage”, “floods”, “warning”, “disasters”, “tsunami”, “natural”, “caused”) addresses natural disasters, and is labeled the same. Others are relatively clear, but require more context to understand. Topic 193 (“human”, “man”, “nature”, “life”, “natural”, “species”, “environment”, “'s”, “biological”, “living”) seems for example to refer to the human species as part of “nature”, but in an ambiguous sense. In this case, we labeled this topic “human species”.

Topic number	Topic label	First 10 words
157	knowledge	systems system based terms nature concept role process forms specific individual
198	experience	s world life time reality human nature form sense man
97	world civilization	world man great human men history today civilization mankind time
140	landscape descriptions	region river area land mountain mountains south north great valley
23	research	study research information studies data scientific work results made carried

[2] Eleven topics feature the word “nature” and another fifteen feature the word “natural”; six of these overlap with the first list.

36	science	science scientific research scientists sciences knowledge technology scientist natural
49	water	water river rivers irrigation dam supply waters dams fresh sea
156	heritage	heritage cultural world monuments sites unesco list site restoration conservation
193	human species	human man nature life natural species environment 's biological living
163	philosophy	philosophy 's philosopher avicenna works thought work knowledge philosophers philosophical
106	national parks	park species conservation biosphere national reserve natural reserves nature world
58	trees	tree trees red green white flowers colour leaves water garden
46	plants	plants plant cells chemical organisms substances cell bacteria micro process
104	climate change	climate environmental change global environment planet carbon world emissions earth
116	forests	forest forests trees soil tropical land tree erosion wood environment
50	natural disasters	earthquake disaster earthquakes damage floods warning disasters tsunami natural caused
20	natural resources	oil iron mining salt copper mineral 's deposits gold coal
91	indigenous peoples	peoples indians indigenous primitive tribes indian people tribe tribal men anthropology civilization
110	sounds	sound noise sounds ear waves hearing vibrations heard vibration noises
102	Darwin/evolution	darwin species whale whales islands galapagos evolution whaling charles natural

Table 1. Nature/natural topics in *Courier* (out of a 200-topic model), in order of strength/weight in the corpus as a whole.

Second, some of the topics are collections of words that are harder to make sense of as a “topic” in the sense of a semantic theme. These suggest rather a type of language, or perhaps genre; a collection of terms that appear together because they are used to articulate a view of or attitude toward a given subject matter. We call these “genre” topics. Topic 157, for example, begins with the words “system” “based” “terms” “nature” “concept” “role” “process” “forms” “specific” “individual”. These terms suggest a metadiscourse about our knowledge of some other topic or topics. We have chosen to call this topic “knowledge systems”, while acknowledging that it is hard to tell what the meaning of this topic could be, in the sense of a semantic cluster. (“Nature” here might for example more often signify “the nature of something”, rather than the natural environment.) Topic 140 (“region” “river” “area” “land” “mountain” “mountains” “south” “north” “great” “valley”), similarly, collects language that might be used to describe landscapes of various types; we label the topic “landscape descriptions”.

Third, we can also observe that the results we get from topic modeling reflect particular features of *Courier* as a publication. For example, most issues of *Courier* were partly or entirely devoted to a particular theme, generally featured on the cover. This feature of the magazine means that a theme that may be rather infrequent in the corpus as a whole is nonetheless distinct in a small number of pages of the magazine (in the relevant special issue or two); this in turn raises the likelihood that such themes will be identified by the topic modeling algorithm as distinct topics. Examples of this are topic 102, labeled "Darwin/evolution" – the subject of a special issue in May 1982 commemorating the 100th anniversary of the scientist's death – and topic 91, regarding indigenous peoples (and labeled as the same). Quite weak in the journal as a whole, this latter topic spikes in likelihood in particular years (as we can see using a tool for measuring topic trends over time). Charting topic 91 then with a tool designed to identify particular pages of *Courier* in which a topic appears most strongly leads us to the thematic issue of summer 1954 entitled, "Last Frontiers of Civilization," which explored "the problem of the world's primitive peoples" [12].

Fourth, we can use the quantitative data from which the topic model is constructed to measure the relative prevalence of these topics in *Courier*. The model's measure of the likelihood of these topic's appearance in the corpus as a whole offers us a quantitative means of ascertaining these topics' relative strength. This shows that some of the broad "genre" topics in which "nature" appears – such as knowledge systems (157), experience (198), or landscape descriptions (140) – are among the very strongest (most likely) topics of all.[3] Because they are so generic, these "topics" appear with many topics (in a narrower, semantic sense of the word) in the corpus. For that reason, they are also of limited utility for our conceptual-historical investigation. Once we exclude these, we can use the same method to measure the relative weight of the remaining topics. Doing this is particularly illuminating if we apply it to the topics as grouped into categories.

We observe namely that the topics – after excluding the "genre" topics – can be divided into three groups. First, we find a set of topics in which nature appears as the subject matter of scientific research, including (in order of relative strength in the corpus): research (23), science (36), human species (193), and Darwin/evolution (102). Second, we see a group of topics in which nature is the object of protection or preservation, including heritage (156), national parks (106), and forests (116). Third, several topics address ways humans live in (and with) nature, in topics such as climate change (104), natural disasters (50), natural resources (20), and indigenous peoples (91). A few topics do not fit this scheme and seem indeed largely irrelevant to our investigation: topic 163 (philosophy), for example, includes the word "nature" but only towards the end of the list of its first fifty words. The distinctions among these groups are of course not water-tight; some topics, including 193 (human species) and 104 (climate change) are intriguing precisely for the way they appear to straddle these categories. It is nonetheless noteworthy that data on topic "weight" shows that the topics in which nature is discussed in the context of scientific research are, as a group, the strongest of the three categories. Those related to preservation come second. The topics we categorize as related to living in nature are weakest.

On the basis of this relatively simple approach, then, we find that *Courier* discussed nature most strongly in terms of science (one of UNESCO's focus areas) and preservation (the heart of UNESCO's commitment to the world's "cultural and natural heritage," which is of course the organization's most famous program). Discussions of nature in the context of "culture"—either in the sense of the arts, or in the sense of particular ways of life—seem by contrast to have been less prominent. But the third group of topics, while "weakest", is also the most difficult to define and, for our purposes, most interesting. So, likewise, are those topics which are harder to place firmly in one of the three categories.

4. Wider contexts of the nature concept through topic networks

Another way to deepen our understanding of the themes surrounding "nature" and "natural" in *Courier* using topic modeling is to study how the topics containing these two terms interacted with other topics in the 200-topic model. Since topic modeling builds on the principle that every topic is represented in every document, although in some cases to a minimal degree, it is possible to construct

[3] Four of these topics (157, 198, 97, and 140) are among the 20 most likely in the entire corpus. The list of all topics from our 200-topic model, ranked by "score", is online at: https://docs.google.com/spreadsheets/d/1-P59S_hvHnXTxsrmnDkEHCimmdW41Bpu/edit?usp=sharing&ouid=101084557857117261252&rtpof=true&sd=true

and visualize networks of co-occurring topics, based on topics that are present on the same magazine page over a given threshold. If studying a single topic says something about that topic's inherent characteristics, examining its broader contexts, or topical associations, can expand our understanding of how, in this case, ideas of nature were envisioned in *Courier*. If a topic about climate change, for example, were to co-occur with a topic about pollution it would generate different connotations than if the climate change topic were to co-occur with one about heritage or art.

In order to study how topics' co-occurrences have changed over time, we compare topic networks for each of our three time periods (1948–1961, 1962–1986 and 1987–2020). To count as a co-occurrence, two topics need to be represented on a *Courier* page at a weight minimum of 0.1 and must co-occur on at least ten pages in the first period and twenty pages in two later periods (in order to take into account the different sizes of the three sub-corpora). Topics that perform below the weight document-score thresholds are thus excluded from the networks. These settings are designed to reveal a network of topic ties that are both representative and strong. For visualization of co-occurring topics, we used the network tool Gephi. The layout algorithm Force Atlas 2 was employed to model the three networks. The sizes of nodes (topic labels) are based on the sum of all individual connections to a topic. The thickness of edges depends on how many documents two topics share. The results are displayed in Figures 1–3, in which topics that contain either "nature" or "natural" (among that topic's first fifty words) are marked with red circles.

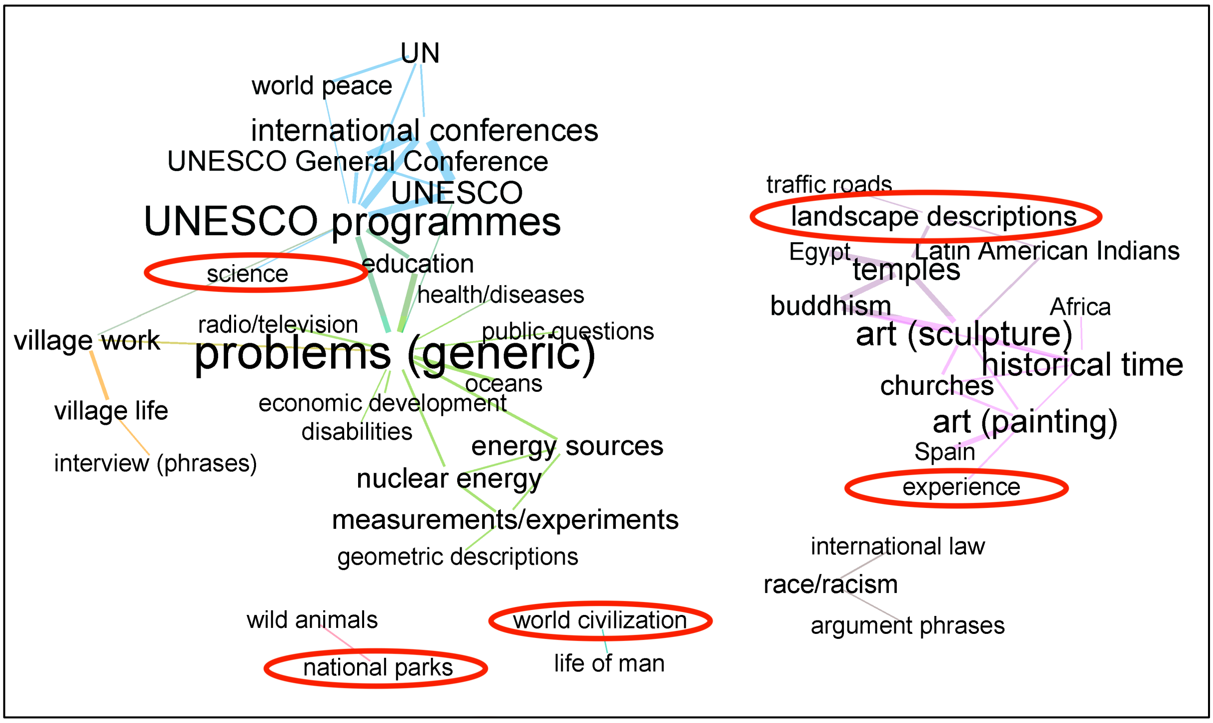


Figure 1. Networks of co-occurring topics for the period 1948–1961. Weight co-occurrence score: 0.1, shared document score: 10. Nature-oriented topics are marked with red circles.

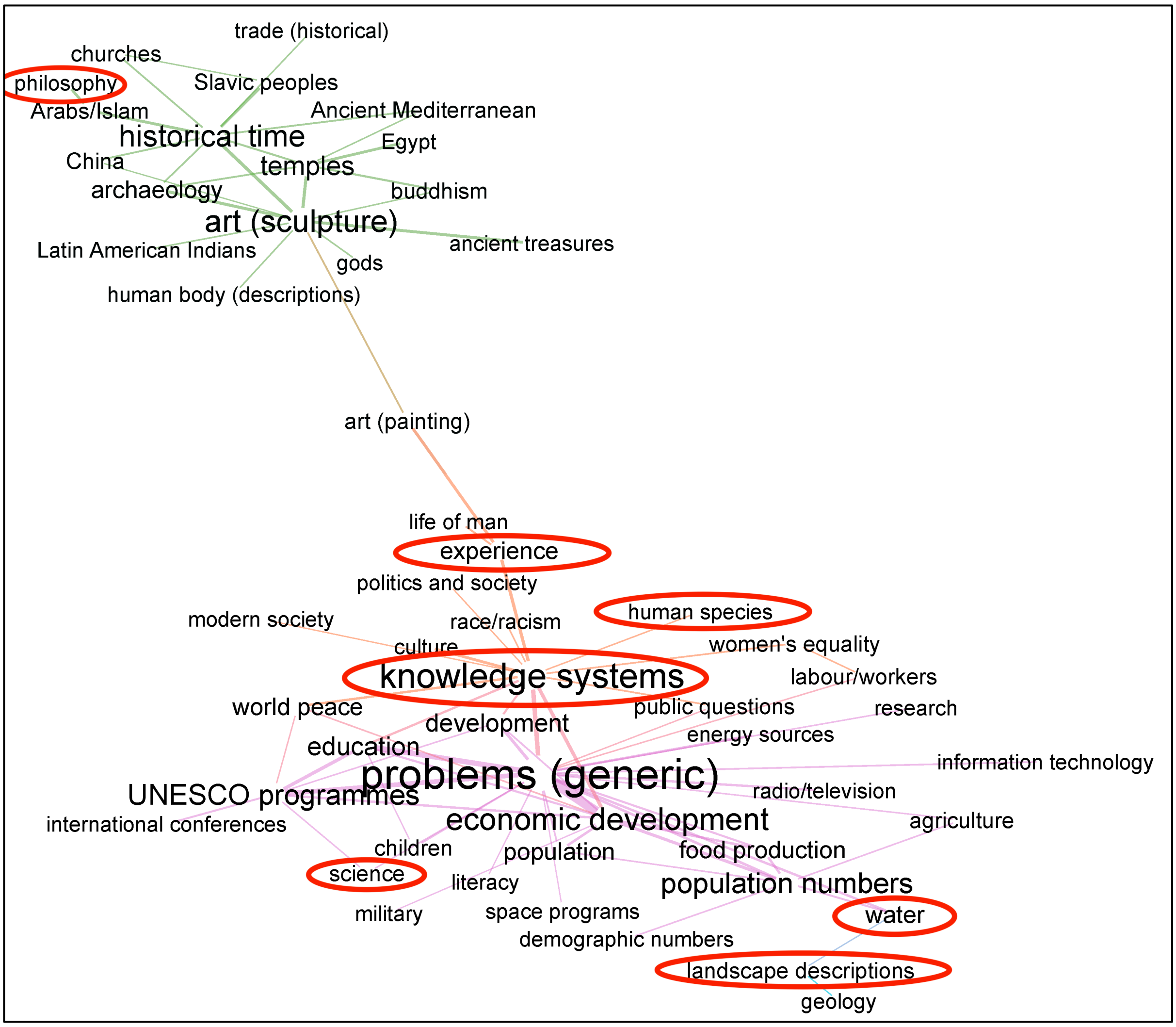


Figure 2. Networks of co-occurring topics for the period 1962–1986. Weight co-occurrence score: 0.1, shared document score: 20. Nature-oriented topics are marked with red circles.

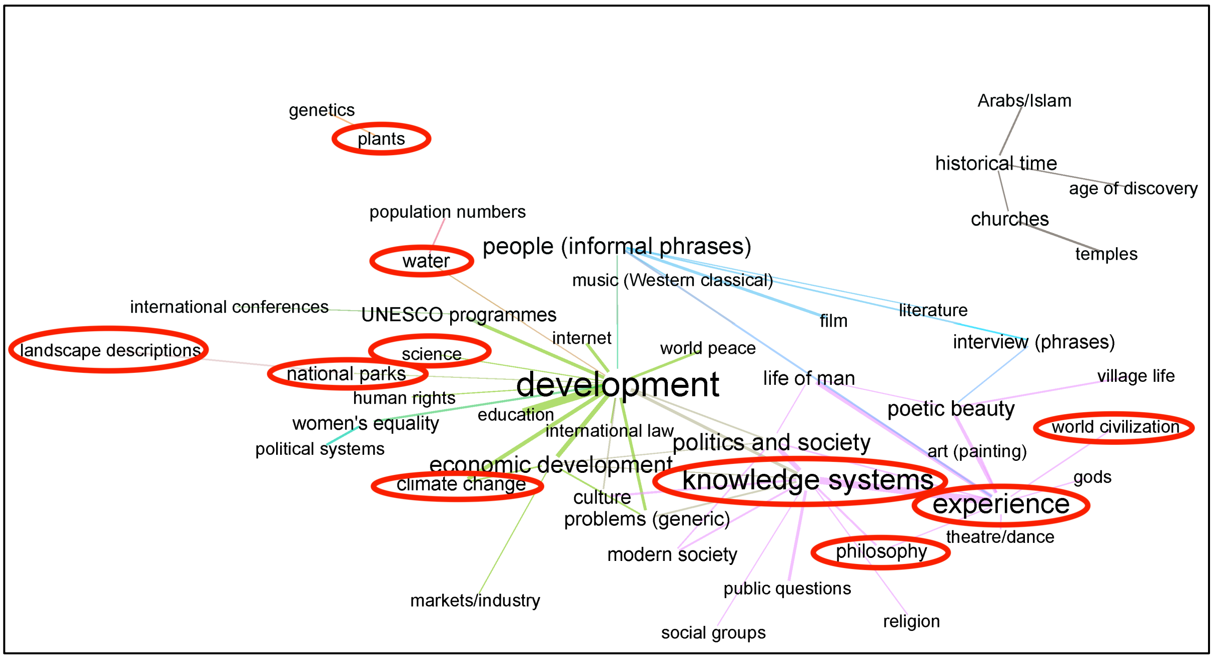


Figure 3. Networks of co-occurring topics for the period 1987–2020. Weight co-occurrence score: 0.1, shared document score: 20. Nature-oriented topics are marked with red circles.

Studying the three networks, some general trends emerge. In each period, a few topics dominate and tie the network together. These include, some of the broad "genre" topics like topic 19 (problems (generic)), topic 157 (knowledge systems), and topic 16 (development). These topics are, as we have seen, among the strongest (by score ranking) in the entire the 200-topic model. Another observation about the network structure concerns how topics are clustered together. The earliest period's network (figure 1) divides into two thematic clusters: one (on the left of the visualization) related to UNESCO itself, including its institutions, programs and two of its three core fields (education and science), and another (on the right) having to do with historical and contemporary cultural and religious expressions. Completely separated in the first period, these two clusters remain essentially distinct in the second period – they are linked only by topic 198 (experience), a "genre" topic that includes terms like "world" "life", "time", "reality", "human" and "nature". In the third period, however, it is difficult to discern clear borders between clusters. This seems like a significant change. What might account for it – a change in the content of *Courier*? An integration of the kinds of generic language use to discuss different topics? A change in the understanding of the relationship between culture and social life? – is impossible to determine from these networks alone. But the change itself raises interesting questions.

Several of the nature-related topics we identified through our first approach do not appear in any of these three networks of co-occurring topics. This does not necessarily mean that, for instance, topic 20 (natural resources), topic 50 (natural disasters), topic 91 (indigenous peoples), topic 116 (forests), or topic 156 (heritage) are not strong themes as such. It simply means that these topics are found to be less likely to co-occur with other *Courier* topics (given the network settings we applied). Indeed, if one excludes the "genre" topics, very few topics connect with more than two other topics. The nature-oriented topic 140 (landscape descriptions) is an interesting exception: it shows up in the first period co-occurring with topic 11 (traffic roads), topic 68 (Latin American Indians) and topic 69 (temples). Topic 140 is also of interest because it is the only nature-oriented topic – excluding "genre" topics and topic 163 (philosophy) – that is situated in the cultural cluster.

Identifying the specific *Courier* pages on which topic 140 (landscape descriptions) co-occurs with other topics in the network offers a means of charting in closer detail how the concept of nature intertwined with discussions of cultural expressions and other human activities. In the period 1948–1961, the topic of landscape descriptions connects to topic 11 (traffic roads) on pages about railroad- and highway-building. "Road teams are taming the wilderness", reads one 1957 article, "stretching the frontiers of civilization, and bringing remote and hitherto untapped areas of the Earth's natural riches under development" [13]. Topic 140 connects also to topic 68 in articles about the way of life of indigenous communities ("Indians") in Latin America. It co-occurs with topic 69 (temples) in articles like a 1955 discussion of St. Catherine's Monastery in Egypt, which is surrounded by mountain peaks: "The savage beauty of this rugged, bare landscape is heightened by the mystery which surrounds the silent rocks" [14]. These examples show how, in the first period, landscape descriptions portrayed nature as wilderness, in contrast to "civilization" and "development", or as the mystical, aestheticized setting of non-Western ways of life. The connection to nature's "savage beauty" and "mystery" is part of what distinguishes these ways of life from modern urban life.

In the second period, 1962–1986, the distinctive words of landscape description appear in the context of natural sites in need of preservation, more or less without human presence, as we see in articles from this period in which topic 140 connects to the topics of water (49) and geology (63), for example in texts about waterways in World Heritage sites and about volcano parks [15]. In the third period, landscape language still links to preservation, but now in a manner that more readily links nature to culture: the pages on which topic 140 connects to topic 106 (national parks) come from articles about parks and nature reserves, in which we find references to the "long and extraordinary interplay between man and nature", and "the aim of harmonizing the needs of people and nature" [16].

On the whole, however, nature-oriented topics appear in the three networks most often in the cluster of topics related to UNESCO and its programs, or to those having to do with development. Nature, in these parts of the networks, appears to be the subject of preservation or intervention for human welfare and development, as we see in figure 3, where "development" is the topic that links topic 49 (water), topic 36 (science), topic 106 (national parks) and topic 104 (climate change).

This form of topic network analysis is helpful, then, in identifying trends in the relationships among contexts in which a particular concept was discussed in the corpus. This mode of analysis can be useful for conceptual history without claiming that the topics themselves represent concepts or discourses. It is enough to accept that the computer-generated topics reflect linguistic patterns in the corpus, some of which are meaningful. We identify our target concept in a quite traditional way, by following its key words (the noun "nature" and the adjective "natural") and using those words' appearance in the computer-generated topics to visualize the relationships among the linguistic contexts in which the concept was invoked, as well as to measure how those contexts and relationships changed over time. This network approach also offers a means of identifying particular passages of interest for close reading, as we have seen. One drawback of the topic network approach is that many topics of interest, like for example topic 193 (human species), do not appear, either because they do not cross the topic weight threshold, or because they were apparently not discussed sufficiently often alongside other (sufficiently strong) topics. Our third approach to topic modeling is an effort to chart such topics, creating a novel means of engaging with the corpus.

5. Using topic modeling to identify thematic paths through the corpus

Here we use the topic model to identify those pages in the magazine in which a chosen topic is particularly strong in a particular time period. Doing this for several of the most intriguing topics produces a set of thematic sequences of pages across time, for manual reading and analysis. Here we explore one example, following topic 193 (human species) through the three time periods designated above. Beginning as it does with the words "human", "man", "nature", "life", "natural", "species" and "environment", this topic offers a compelling way to chart historically the uses of the concept of nature (expressed through the noun "nature" and the adjective "natural") in the corpus. Moreover, the topic's inclusion of the words "human" and "man" suggests an interest in the relationship with the realm of culture, broadly construed. Following one computer-generated topic will not of course reflect all uses of the concept. But that is the strength of the approach: rather than giving us a generic sample of places where the word appeared, this topic points us toward one specific linguistic context in which our target words were invoked.

The topic modeling algorithm we employ here enables us to identify the pages (in a given time period) on which a selected topic was strongest. (The relevant measure is of "topic weight", a value that is normalized to the number of words on a given page.) We selected the first three pages (per period) for manual reading. What we found by reading articles in which topic 193 was strongest was a striking set of changes over time in the relationship between the concepts of nature and culture across our selected time periods.

In *Courier*'s early years (1948–1961), the articles in which topic 193 was especially strong emphasized a sharp distinction between nature and culture. These articles suggest that authors writing in *Courier* did this in part because that distinction was the core of an antiracist vision of humanity, inspired by contemporary trends in anthropology. A central point of post-1945 international antiracism was to argue that what accounted for differences among human groups, in terms of their ways of life or their apparent differences in social or intellectual aptitude, was not nature (biology, genetics), but culture. We see these arguments in the earliest articles highlighted by following topic 193, like a 1953 article by the Rutgers University anthropologist Ashley Montagu on the idea of "human nature". Here Montagu argues that man is "the most plastic, the most malleable, the most educable, of all living creatures" because our real differences are not a matter of nature, just culture. "Science", Montagu writes, "knows of no *natural* drive in man to make knives and forks or to speak Italian; Australian aborigines neither use knives or forks nor do they speak Italian, not because they couldn't do so, but because they happen to be born into a *cultural* environment from which such instruments are absent and where their own language alone is spoken" [17]. Similar invocations of a strong dichotomy between nature and culture can be found early in the following period (1962–1986). There, topic 193 leads us for example to a 1965 article by the Belgian researcher Jean Hiernaux on "cultural evolution" – namely, the progress that humankind has made and makes by accumulating and passing down knowledge and skills, not through genetic change. Here, too, we find the concepts of nature and culture discussed in close relationship to one another, and again, more or less as opposites [18].

In the second period (1962–1986), we see signs of a different nature discourse, as topic 193 leads us to articles arguing that nature and culture are (or should be) in a delicate balance – not opposites, but integrated. Topic 193's strongest appearances in period two include a 1969 article by René Dubos, in which the French-born American microbiologist and prominent public intellectual calls for a new mix of human agency (culture) and the physical, non-human environment (nature), ideally resulting in what he refers to as "civilized nature". Motivating this call for a balanced integration of nature and culture is Dubos's doubt that the two spheres can really be understood in isolation from one another. We see this where he comments explicitly on the weakness of the prevailing understanding of "nature": "The ill-defined meaning of the word Nature compounds the difficulty of formulating a scientific basis for the philosophy of conservation. If we mean by nature the environment as it would exist in the absence of man, then very little of it survives" [19]. Nature wholly distinct from culture is conceivable to Dubos, but so rare as to be almost irrelevant.

In the third period, beginning in 1987, topic 193 identifies discussions in which new ways of thinking about nature are pushing into areas that were once purely the domain of culture. A 1993 interview with the French philosopher Luc Ferry discusses the notion that nature might have rights, and the question "to what extent does ecology question the basis of modern civilization" [20]. Articles like this one gave voice to doubts about the possibility or desirability of what Dubos had called "civilized nature". Now there is concern that nature must be understood to have systems of its own (embraced by the newly popular category of ecology) as well as values or even rights of its own, apart from any human cultural viewpoint.

But the most recent articles identified by following topic 193 show still another turn. This same topic leads us to a 2018 interview in which the historian Dipesh Chakrabarty introduced *Courier*'s readers to the concept of the Anthropocene, the name recently suggested for our current geological age, characterized by humankind's impact on planetary systems, including above all the climate. Here Chakrabarty argues that "humans are a geological force", such that nature and culture cannot be separated even at the level of the planet [21].

This brief tour of *Courier* through topic 193 shows us that Chakrabarty's point was not so new: *Courier* authors had suggested the interconnection of culture and nature back in the 1960s, albeit in less radical terms. We see, moreover, that what was at stake in arguing for or against a sharp nature-culture distinction changed a great deal over the last half-century. In the 1950s, progressive voices at UNESCO insisted on a sharp nature-culture distinction as part of the fight against racism and for a new world order of human equality. By the early twenty-first century, it seemed like a matter of great urgency that we abandon a sharp nature-culture distinction, so that we could understand and act on the existential threat to humanity posed by climate breakdown.

This was just one of multiple such explorations of the corpus that this approach permits. It is an exploration that is, ultimately, old-fashioned in its focus on reading and contextualizing. But it was driven by a combination of word-searching with algorithmically generated topics, and would not have worked otherwise.

6. Using topic modeling to identify thematic paths through the corpus

The research documented in this article leads us to two conclusions, one methodological and one empirical. Regarding methodology: we have explored three modes of using topic modeling for conceptual history. The selected approaches seem to work better, or at any rate differently, at different scales. Each scale offers a different perspective on the corpus, from the most distant (topic networks) to the closest (reading pages identified by following topics of relevance to a chosen concept). We find, too, that the approaches complement one another. Observations we make by reading passages in which a topic is particularly strong can be put into perspective by the quantitative data that comes from having modeled the entire corpus. We know, for example, that topic 193, while interesting, was not terribly strong in the corpus as a whole (in terms of topic weight it ranks at place 83 out of 200). On the other hand, our findings from close reading articles signaled by this topic offer perspective on the findings we made through the more “distant” approaches used earlier in the paper, in the sense that we find that a nature topic that seemed to be focused on nature as the object of scientific research was, in fact, a highly significant context for discussions of the nature-culture relationship.

As for an empirical conclusion: the concepts of nature and culture engaged in a complex dance in *Courier*, now closer, now further apart, and the journal seems to offer rich resources in which to follow it. Judging from the results obtained by following topic 193, the dance seems to have been led by European thinkers, with the important exception of the Indian-born historian Dipesh Chakrabarty. Does this mean that the global character of the magazine did not amount to much, as far as this theme is concerned? That will be an interesting question to explore. We are encouraged that topic modeling will offer one stimulating means of doing so.

Acknowledgments

Research for this paper was supported by a grant from the Swedish Research Council for the project “International Ideas at UNESCO: Digital Approaches to Global Conceptual History” (VR 2019-03278). Thanks to the UNESCO Archives, and especially archivist Eng Sengsavang, for facilitating access to document. Many thanks also to Roger Mähler and Andreas Marklund of Humlab (Umeå University) for their invaluable research support.

References

[1] A selection of articles is available in *UNESCO Courier: Transforming Ideas. Selected Articles*, Volume I: *Thinkers*, and Volume II: *Creators* (Paris: UNESCO, 2021).

[2] See for example “Culture” and “Nature” in Williams, Raymond. *Keywords: A Vocabulary of Culture and Society* (Oxford: Oxford University Press, 1986).

[3] On the importance of the post-1945 period for the environment, and for the emergence of the modern concept of the environment, see: McNeill, J. R., and Peter Engelke. *The Great Acceleration: An Environmental History of the Anthropocene since 1945* (Cambridge, MA: Belknap Press of Harvard University Press, 2016); Robin, Libby, Sverker Sörlin, and Paul Warde. *The Environment: A History of the Idea* (Baltimore: Johns Hopkins University Press, 2018).

[4] Defourny, Vincent. “Public Information in the UNESCO”, in *The Global Public Relations Handbook: Theory, Research, and Practice*, edited by Krishnamurthy Sriramesh and Dejan Verčič, (Mahwah, NJ: Lawrence Erlbaum, 2003), 428-429. On *Courier’*s history, see also: Simonsen, Maria, “Routes of Knowledge: The Transformation and Circulation of Knowledge in the UNESCO Courier, 1947-1955”, in *Forms of Knowledge: Developing the History of Knowledge*, edited by Johan Östling, David Larsson Heidenblad, and Anna Nilsson Hammar (Lund: Nordic Academic Press, 2020); Krebs, Edgardo C. “Popularizing Anthropology, Combating Racism: Alfred Métraux at The UNESCO Courier”, in *A History of UNESCO: Global Actions and Impacts*, edited by Poul Duedahl (Houndmills: Palgrave Macmillan, 2016), 29–48.

[5] Blei, David M. “Probabilistic Topic Models.” *Commun. ACM* 55, no. 4 (April 2012): 77–84. https://doi.org/10.1145/2133806.2133826; for Mallet see: https://mimno.github.io/Mallet/

[6] Mohr, John W., and Petko Bogdanov. “Introduction—Topic Models: What They Are and Why They Matter.” *Poetics*, Topic Models and the Cultural Sciences, 41, no. 6 (December 1, 2013): 545–69. https://doi.org/10.1016/j.poetic.2013.10.001, here 564.

[7] Maurel, Chloé. *Histoire de l’Unesco: les trente premières années, 1945-1974* (Paris: L’Harmattan, 2010).

[8] Brouillette, Sarah. *UNESCO and the Fate of the Literary* (Stanford, CA: Stanford University Press, 2019), 2; Weber, Raymond. “Les Organisations multilatérales face aux nouveaux défis de la coopération culturelle.” In *Géopolitique de la culture: Espace d’identité, projections, coopération*, edited by François Roche (Paris: Harmattan, 2007), 83.

[9] On transnational and global conceptual history see for example: Moyn, Samuel, and Andrew Sartori, eds. *Global Intellectual History* (New York: Columbia University Press, 2013); Pernau, Margrit, and Dominic Sachsenmaier, eds. *Global Conceptual History: A Reader* (London: Bloomsbury Publishing, 2016); Jordheim, Helge, and Erling Sandmo, eds. *Conceptualizing the World: An Exploration across Disciplines* (New York: Berghahn Books, 2018.)

[10] Boyden, Michael, Ali Basirat, and Karl Berglund. “Digital Conceptual History and the Emergence of a Globalized Climate Imaginary.” *Contributions to the History of Concepts* 17, no. 2 (December 2022): 95–122: De Bolla, Peter, Ewan Jones, Paul Nulty, Gabriel Recchia, and John Regan. “Distributional Concept Analysis.” *Contributions to the History of Concepts* 14, no. 1 (June 2019): 66–92; Gavin, Michael, Colin Jennings, Lauren Kersey, and Brad Pasanek. “Spaces of Meaning: Conceptual History, Vector Semantics, and Close Reading.” In *Debates in the Digital Humanities 2019*, edited by Matthew K. Gold and Lauren F. Klein (Cambridge, MA: MIT Press, 2019), 243–67.

[11] Mohr and Bogdanov, “Topic models”, 561.

[12] Alfred Métraux, “Editorial: The Problem of the World’s Primitive Peoples,” *Courier* 7, no. 8–9 (1954), 3–4.

[13] W. H. Owens, “Towards a World Highway”, *Courier* (June 1957), 4.

[14] Albert Raccah, “In Mt. Sinai Desert: The Monastery of the Burning Bush”, *Courier* (March 1955), 22.

[15] Harold J. Plenderleith, Caesar Voute, and Theodoor de Beaufort, “Mohenjo-Daro: A 5,000-Year Old Heritage Menaced by Destruction”, *Courier* (June 1965); Haroun Tazieff, “Confronting the Irascible Irazu”, *Courier* (November 1965).

[16] “Return of the Griffon”, *Courier* (October 1987), 14; Alison Jolly, “The World’s Wild Places”, *Courier* (August 1988), 6.

[17] Ashley Montagu, “Human Nature Cannot Be Changed: False, Says U.S. Anthropologist”, *Courier* 6, 2 (February 1953), 15.

[18] Jean Hiernaux, “The Future of ‘Homo Sapiens’, *Courier* 18, 4 (April 1965), 12–15.

[19] René Dubos, “The Biosphere: A Delicate Balance Between Man and Nature”, *Courier* 22, 1 (January 1969), 7–16; quotations from 12.

[20] “Interview: Luc Ferry talks to Bahgat Elnadi and Adel Rifaat”, *Courier* 46, 4 (April 1993), 4.

[21] Dipesh Chakrabarty, interviewed by Shiraz Sidhva, “Humans are a geological force”, *Courier* (April-June 2018), 11–14.

An OCR Pipeline for Transforming Parliamentary Debates into Linked Data: Case ParliamentSampo - Parliament of Finland on the Semantic Web

Senka Drobac[1], Laura Sinikallio[1,2] and Eero Hyvönen[1,2]

[1]Aalto University (Semantic Computing Research Group (SeCo)), Finland
[2]University of Helsinki (HSSH, HELDIG, Cultural heritage studies), Finland

Abstract

This paper presents the OCR pipeline created for *ParliamentSampo - Parliament of Finland on the Semantic Web*, a Linked Open Data (LOD) service, data infrastructure, and semantic portal for studying Finnish political culture, language, and networks of the Members of Parliament (MP). A knowledge graph of linked data has been created based on ca. 967 000 speeches in all plenary sessions of the Parliament of Finland in 1907—2022; the data is also available in XML format, utilizing the new international Parla-CLARIN format. A central part of the historical debates 1907-1999 was available only as PDF documents of fairly low OCR quality and had to be OCRed first; this paper reports lessons learned from this process.

1. Introduction

Parliamentary data are used in many areas of research [1], as they provide a wealth of information about the state and functioning of democratic systems, political life and, more generally, language and culture. The most prominent part of the work of parliaments is the public plenary sessions, in which the Members of Parliament (MP) discuss and vote on issues on the agenda and other topics that arise [2]. Semantic Web (SW) technologies[1] and Linked Data (LD) [3] provide a promising approach for publishing and using parliamentary data in Digital Humanities (DH) [4, 5, 2]. The LD approach for Cultural Heritage [6] has arguably many advantages, including:

- Linked data and ontologies [7] provide a framework for harmonizing heterogeneous distributed datasets and combining them into larger and richer entities.
- The SW is based on the Predicate Logic [8], which provides an opportunity to enrich data by linking new information.
- When machines can "understand" data content, intelligent web services and data analyses can be implemented more easily.

DHNB2023 | Sustainability: Environment - Community - Data. The 7th Digital Humanities in the Nordic and Baltic Countries Conference. Oslo – Stavanger – Bergen, Norway. March 8–10, 2023.

senka.drobac@aalto.fi (S. Drobac)
0000-0002-7645-3079 (S. Drobac); 0000-0003-1695-5840 (L. Sinikallio); 0000-0003-1695-5840 (E. Hyvönen)

DHNB Publications, DHNB2023 Conference Proceedings, https://journals.uio.no/dhnbpub/issue/view/875

[1] https://www.w3.org/standards/semanticweb/

- Ready-made tools by other actors can be re-used for publishing, processing and analysing the standardized data.

However, using linked data requires that the typically textual, unstructured debates have to be transformed into semantic structured data in several steps:

1. If the minutes are available only in print they have to be first digitized.
2. Texts have to be OCRed from digitized documents.
3. Metadata about the OCRed texts has to be extracted and represented using RDF.[2]
4. The data can be enriched and interlinked and finally be published and made available in a SPARQL endpoint.
5. Applications on top of the endpoint can be created or the data service can be used for data-analytic research.

This paper concerns step 2 in the case of publishing and using Finnish parliamentary speech data. In this case, the digitized data was provided by the open data service of the Parliament of Finland (PoF).[3] Metadata extraction and enrichment (steps 3-4) are described in [9, 10, 2]. The speech data outcome described in this paper has already been used as a basis for analyzing concepts in political speeches [11], for network analyses based on MP references in speeches [12], and for data analyses of speeches and for portal application development [2].

2. Related Work

The minutes of parliamentary plenary debates have been compiled into several corpora, allowing for analysis of their content and language (e.g., the corpus of the Norwegian parliament [13]; CLARIN list of parliamentary corpora[4]). These corpora are represented using the TEI-based Parla-CLARIN scheme[5], which has been developed within the CLARIN infrastructure to provide a unified standard [14]. The related ParlaMint project[6] is dedicated to creating comparable national parliamentary corpora based on the Parla-CLARIN scheme. Additionally, parliamentary materials have been transformed into Linked Data format for the creation of systems such as LinkedEP [4], which deals with European Parliament data, as well as the Italian Parliament[7] and LinkedSaeima for the Latvian parliament [5].

The materials of PoF have been digitized in various contexts but are difficult to use, as they have been produced separately from different periods and stored in different formats [9]. The usability of the materials is also hampered by their varying quality and lack of descriptive data [15]. Language corpora have been published on parliamentary debates, such as the Parliamentary Corpus of FIN-CLARIN's Language Bank[8] [16] which covers the years 2008-2016. It

[2] https://www.w3.org/RDF/

[3] https://avoindata.eduskunta.fi/#/fi/digitoidut/

[4] https://www.clarin.eu/resource-families/parliamentary-corpora

[5] https://github.com/clarin-eric/parla-clarin

[6] https://www.clarin.eu/content/parlamint-towards-comparable-parliamentary-corpora

[7] http://data.camera.it

[8] http://korp.csc.fi

contains the speeches in a linguistically annotated form and also synchronized links to original plenary session videos [17]. The Voices of Democracy project has produced a research corpus that includes plenary minutes in 1980-2018 annotated grammatically as well as interviews of veteran MPs conducted by the PoF after 1988 [18]. The minutes of the parliamentary debates from 1991 to 2015 can also be found in the International Harvard Parlspeech Corpus [19], but we have identified gaps in the coverage in this corpus.

Some of the most popular open source OCR tools for historical printed texts are Tesseract[9], OCR4All [20] and OCR-D [21]. In Finland, the most OCR efforts have been focused on newspaper material [22] and [23]. A comprehensive post-correction survey is presented in [24].

3. The OCR Pipeline

This section presents the OCR pipeline used in transforming the Finnish debate corpus 1907-1999 into LD.

Data sources The parliamentary speech data is provided in three different file formats. Parliamentary sessions from the period 1907-1999 have been scanned and made available as PDF files. Later data is already in machine-readable formats, with sessions from 1999-2015 in HTML and from 2015 onward in XML format.

The performance of OCR systems is heavily influenced by the quality of the input data. In our situation, the PDF documents are generally of high quality, which simplifies the OCR process. However, data from the early 1920s is considerably noisy due to the use of low-quality paper during that era, as illustrated in Figure 1. Moreover, there are occasional skewed pages that present difficulties for OCR. All the data is printed in contemporary Latin fonts, with minor variations over time. As for formatting, early minutes are presented in a single-column layout, while the majority of the data is presented in a double-column format. In earlier documents, the double columns were separated by a black line, whereas in later ones, they were separated by white space.

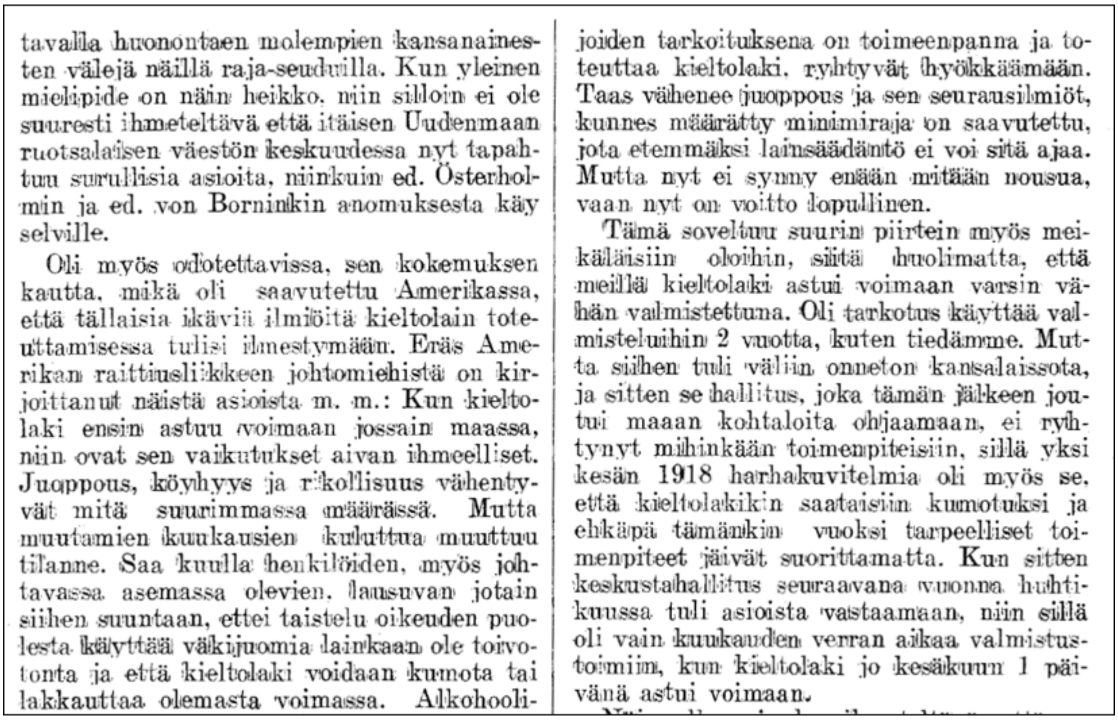
The OCR Process We OCRed the data with Tesseract 4, with pre-trained models for both Finnish and Swedish. We opted for Tesseract due to its high accuracy pre-trained models, capable of recognizing various contemporary fonts, and its ability to use multiple language models concurrently.

We were able to perform OCR quickly by omitting the training stage, and the ability to support multiple languages was particularly crucial since parliamentary speeches in Finland are primarily given in two official languages: Finnish and Swedish. Fig. 2 illustrates the proportion of Finnish and Swedish languages used in these speeches over time. Finnish is the predominant language used, and the percentage of Swedish has gradually decreased over time, from 18-20% to minimal levels today.

To begin the OCR process, we first converted the PDF files into images. Tesseract's documentation[10] suggests that the software performs optimally on images with a minimum DPI of 300. After conducting preliminary tests on a small dataset with various image resolutions, we discovered that the best OCR outcomes were obtained at a resolution of 350 DPI. Curiously, both

[9] https://github.com/tesseract-ocr/tesseract

[10] https://github.com/tesseract-ocr/tessdoc/blob/main/ImproveQuality.md



tavalla huonontaen molempien kansanainesten välejä näillä raja-seuduilla. Kun yleinen mielipide on näin heikko, niin silloin ei ole suuresti ihmeteltävä että itäisen Uudenmaan ruotsalaisen väestön keskuudessa nyt tapahtuu surullisia asioita, niinkuin ed. Österholmin ja ed. von Borninkin anomuksesta käy selville.

Oli myös odotettavissa, sen kokemuksen kautta, mikä oli saavutettu Amerikassa, että tällaisia ikäviä ilmiöitä kieltolain toteuttamisessa tulisi ilmestymään. Eräs Amerikan raittiusliikkeen johtomiehistä on kirjoittanut näistä asioista m. m.: Kun kieltolaki ensin astuu voimaan jossain maassa, niin ovat sen vaikutukset aivan ihmeelliset. Juoppous, köyhyys ja rikollisuus vähentyvät mitä suurimmassa määrässä. Mutta muutamien kuukausien kuluttua muuttuu tilanne. Saa kuulla henkilöiden, myös johtavassa asemassa olevien, lausuvan jotain siihen suuntaan, ettei taistelu oikeuden puolesta käyttää väkijuomia lainkaan ole toivotonta ja että kieltolaki voidaan kumota tai lakkauttaa olemasta voimassa. Alkohooli-

joiden tarkoituksena on toimeenpanna ja toteuttaa kieltolaki, ryhtyvät hyökkäämään. Taas vähenee juoppous ja sen seurausilmiöt, kunnes määrätty minimiraja on saavutettu, jota etemmäksi lainsäädäntö ei voi sitä ajaa. Mutta nyt ei synny enään mitään nousua, vaan nyt on voitto lopullinen.

Tämä soveltuu suurin piirtein myös meikäläisiin oloihin, siitä huolimatta, että meillä kieltolaki astui voimaan varsin vähän valmistettuna. Oli tarkotus käyttää valmisteluihin 2 vuotta, kuten tiedämme. Mutta siihen tuli väliin onneton kansalaissota, ja sitten se hallitus, joka tämän jälkeen joutui maaan kohtaloita ohjaamaan, ei ryhtynyt mihinkään toimenpiteisiin, sillä yksi kesän 1918 harhakuvitelmia oli myös se, että kieltolakikin saataisiin kumotuksi ja ehkäpä tämänkin vuoksi tarpeelliset toimenpiteet jäivät suorittamatta. Kun sitten keskustahallitus seuraavana vuonna huhtikuussa tuli asioista vastaamaan, niin sillä oli vain kuukauden verran aikaa valmistustoimiin, kun kieltolaki jo kesäkuun 1 päivänä astui voimaan.

Figure 1: A snippet from a 1921 document shows smudged text due to the poor quality of the paper.

lower and higher image resolutions yielded inferior OCR results. After preparing the images, we performed the recognition with Tesseract, using `-l fin+swe` option, which prioritizes Finnish as the language for recognition while also having the capability to recognize Swedish.

Due to the extensive size of our dataset, consisting of 324,333 page images, we used CSC's supercomputer Puhti[11] provided by CSC - IT Center for Science[12] for OCR. This supercomputer enabled us to leverage GPU computing and perform up to 100 SLURM batch jobs in parallel using its *array jobs* feature. Once we set up the parallel system (with an average of around 3,200 pages recognized per GPU at a time), the OCR process took only a few hours (typically between 5 to 8 hours, depending on the job size). Accounting for the queuing time for the resources, the entire recognition process was completed in a matter of days.

Post-correction and Transformation into Linked Data and Parla-CLARIN

To gather all speeches of the Finnish Parliament in the 20th century, we used pattern recognition and regular expressions on the plain-text version of the OCR results. The OCRed results were satisfactory as they were, but to enhance the reliability of the gathered data we performed a few manual corrections to the OCR results. Each transcript of a plenary session started with a title row that spanned the whole page, whereas the rest of the document was mainly split into two columns. Due to this, the title was sometimes split into two rows or otherwise corrupted

[11] https://research.csc.fi/-/puhti

[12] https://www.csc.fi/en/

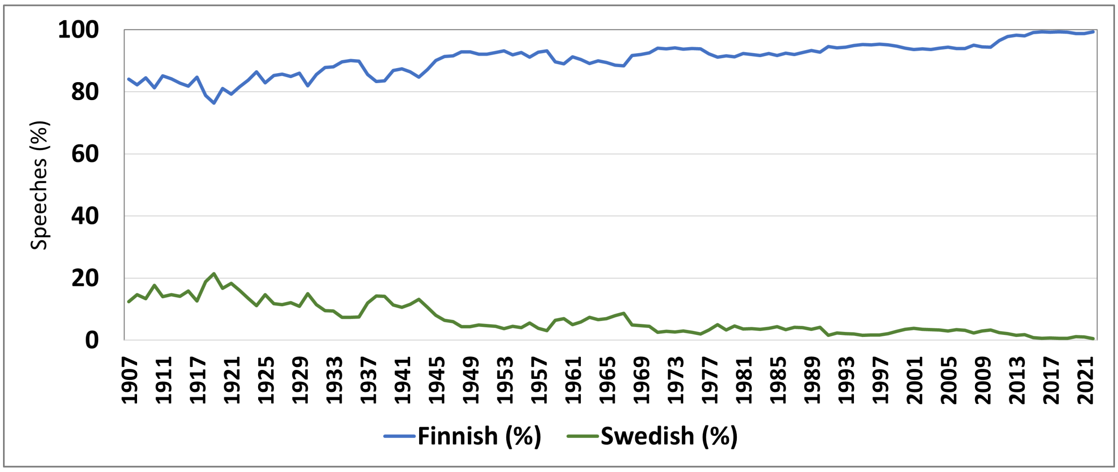


Figure 2: The graph shows percentages of language representation in speeches through years. The blue line shows the percentage of Finnish text and the green line the percentage of Swedish data. The graph has been calculated on a sentence level.

in the results. As one file included several transcripts, one after another, these title rows and the information they contained (date, session number) were central in connecting speeches to correct sessions. With a helper script, all distorted title rows were located and manually corrected.

After corrections, we created Python scripts to scrape all relevant data from the OCRed text files. First, we gathered speeches and their metadata in CSV format. Then the data went through several automated correction and enrichment steps. A central part of the enrichment was retrieving speaker information from an external *Members of Parliament* (MP) dataset [10] as the transcripts contained only each speaker's title and surname. The correct person was found based on the scraped surname and session date only, so for correct linking, the names needed to be correct. Hence the majority of the correction efforts went into fixing speaker names and titles that had been distorted in the OCR process (e.g. *Procopé* had become *Procop&*).

Typical correction steps were: Handling of missing or extra whitespaces (*MinisteriHuttu* → *Ministeri Huttu, Ministeri Lin n ain maa* → *Ministeri Linnainmaa*), removal of extra trailing characters, such as special characters, and replacing some systematically recurring errors, such as a common surname ending *qvist* having become *gvist*. If the corrections weren't enough to find the right match, we would pick the closest match from the list of all possible surnames from the MP data set.

Finally, we transformed the speeches into two parallel data sets: (1) an RDF (*Resource Description Framework*) [25] format speech knowledge graph, forming linked data and (2) an XML corpus formed according to the Parla-CLARIN v0.2 specification [26]. More on this transformation can be read from [9].

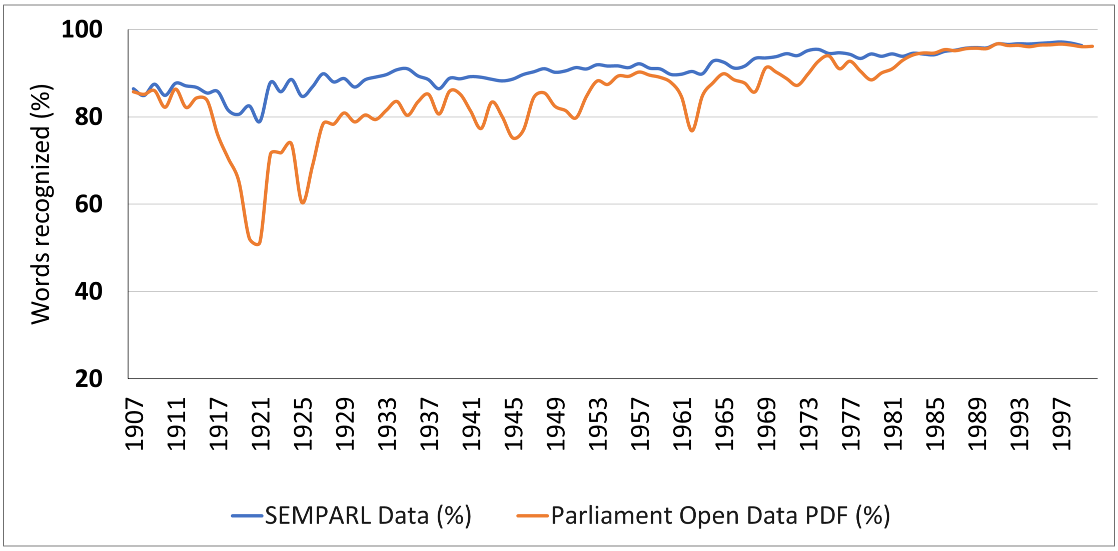


Figure 3: The percentage of recognized words with LAS tool on PoF OCR results (orange) and our new OCR (blue) results.

4. Evaluation

In order to evaluate the final results, we utilized the Language Analysis Command-Line Tool (LAS) [27] to calculate the percentage of correctly recognized words. LAS has been specifically adapted to work well for Finnish and uses Finite State Machines to verify the presence of words in Finnish morphological lexical database Omorfi [28].

This tool is particularly useful for a task like this because it covers all possible grammatical word forms in a language. However, a limitation of such a tool is the dictionary's scope, particularly with regards to specialized and historical language. Although the LAS tool has expanded the original Finnish morphology to cover historical spelling variations, it has not been adapted to legal vocabulary.

In the first experiment, we conducted a comparison between the accuracy of our OCR texts and those provided in the PoF original documents. Figure 3 shows the percenteges of recognized words with LAS tool on the received material (orange line) and our results (blue line). This evaluation was carried out on the entire dataset, not just speeches, and using only Finnish morphology. Recognizing text in complete documents is generally more challenging than in speeches, because they may contain structures other than running text (eg., tables and lists). Additionally, the dataset is bilingual, resulting in lower overall accuracy. Nevertheless, this experiment provides us with an indicator of the improvement we are pleased to report in OCR accuracy, particularly during the early 1920s when the dataset was most challenging

In addition to the evaluation on the entire dataset, we also conducted a separate evaluation solely for speeches. However, since the tool can only use one language for a given string and many speeches, particularly in the early years, were bilingual, we tokenized speeches

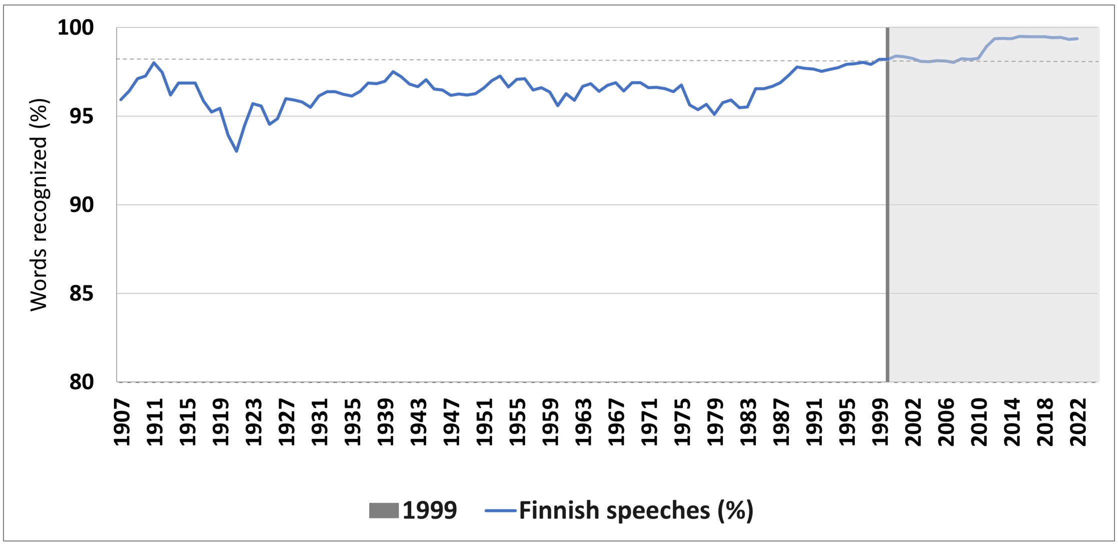


Figure 4: The percentage of recognized words with LAS tool on Finnish speeches (blue line). Results prior to 1999 indicate the performance on OCR text, where those after were based from native digital data.

into sentences using Python's `nltk.tokenize`[13] and performed recognition on a sentence level. Conveniently, the tool also performs language recognition, making this approach both convenient and effective.

The results of the evaluation on the Finnish speeches are presented in Figure 4. The blue line represents the accuracy of the Finnish data. A vertical line denotes the year 1999, which signifies the point before the data was OCRed and after which it was available in HTML and XML formats. This information helps in evaluating the quality and scope of the morphology employed.

The graph demonstrates that the percentage of correctly recognized words remains consistently above 95%, except for the period around 1920 when the scanned images were particularly noisy. On the right side of the graph, the results for natively digital data illustrate that the benchmark in the early 2000s is approximately 98%. This indicates that the accuracy of the OCR data is only slightly lower, ranging from 0-3% below the benchmark.

Similarly, we conducted an evaluation on Swedish speeches; however, the benchmark was low, with recognition rates mostly ranging from 86% to 93%. The OCR accuracy rates were mostly between 85% and 92%, with early data up to 1916 showing recognition rates of 80-82%.

5. Discussion and Conclusion

Although higher resolution images are often thought to yield better OCR results, we discovered that in our case, using resolutions greater than 350 dpi led to poorer accuracy. This could be

[13] https://www.nltk.org/api/nltk.tokenize.html

because the pre-trained models we used were developed using images with that resolution.

The most practical way we found to evaluate the data was to use the LAS tool since we did not have any Ground Truth (GT) data. Creating GT data would have been a time-consuming process, especially considering the variations throughout the years. We would have to create GT data for each year, and to obtain a good evaluation, we would need to sample a significant amount of data from every year. It would not be feasible to capture entire pages, so we would need to select lines from different pages, create GT and evaluate against the produced OCR. This would require annotating at least 150-200 lines from each year, totaling around 13,800-18,000 lines. It is uncertain whether this effort would yield more information on data quality, especially for Finnish text. Our approach provided us with a reasonable understanding of OCR quality, despite not having good conclusive quality estimation for the Swedish proportion of the corpus.

The results with LAS tool on the Swedish speeches are difficult to interpret since the low benchmark makes it unclear whether the unrecognized words come from the the OCR errors or the small scope of the morphological acceptor. The majority of the data falls within the benchmark range, but since the range is so large, it is impossible to draw any definitive conclusions based solely on these results.

The Finnish OCR quality is outstanding, with only a 0-3% difference from the benchmark in terms of the number of recognized words, except for the early 1920s period where lower accuracy was due to poor image quality. One potential solution for future work could be to train recognition models on noisy data from that period, or explore the extent of improvement that could be achieved through fine-tuning existing models. Furthermore, to avoid the need for manual post-correction of titles, we could experiment with using an XML version of the OCR results. It includes text coordinates in images and it could enable us to automatically identify split titles. This would make the entire process fully automatic and easily reproducible.

Acknowledgments

Our work is funded by the Academy of Finland and is also related to the EU project InTaVia[14] and the EU COST action Nexus Linguarum[15]. We used the computing resources of the CSC - IT Center for Science.

References

[1] C. Benoît, O. Rozenberg (Eds.), Handbook of Parliamentary Studies: Interdisciplinary Approaches to Legislatures, Edward Elgar Publishing, 2020. doi:10.4337/9781789906516.

[2] E. Hyvönen, L. Sinikallio, P. Leskinen, M. L. Mela, J. Tuominen, K. Elo, S. Drobac, M. Koho, E. Ikkala, M. Tamper, R. Leal, J. Kesäniemi, Finnish parliament on the semantic web: Using ParliamentSampo data service and semantic portal for studying political culture and language, in: Digital Parliamentary data in Action (DiPaDa 2022), Workshop at the 6th Digital Humanities in Nordic and Baltic Countries Conference, long paper, CEUR Workshop Proceedings, Vol. 3133, 2022. URL: http://ceur-ws.org/Vol-3133/paper05.pdf.

[14] https://intavia.eu

[15] https://nexuslinguarum.eu

[3] T. Heath, C. Bizer, Linked Data: Evolving the Web into a Global Data Space (1st edition), Synthesis Lectures on the Semantic Web: Theory and Technology, Morgan & Claypool, 2011. URL: http://linkeddatabook.com/editions/1.0/.

[4] A. Van Aggelen, L. Hollink, M. Kemman, M. Kleppe, H. Beunders, The debates of the European Parliament as Linked Open Data, Semantic Web – Interoperability, Usability, Applicability 8 (2017) 271–281. doi:10.1007/s42001-019-00060-w.

[5] U. Bojārs, R. Darģis, U. Lavrinovičs, P. Paikens, LinkedSaeima: A linked open dataset of Latvia's parliamentary debates, in: Semantic Systems. The Power of AI and Knowledge Graphs. SEMANTiCS 2019, Springer, 2019, pp. 50–56. doi:10.1007/978-3-030-33220-4_4.

[6] E. Hyvönen, Publishing and Using Cultural Heritage Linked Data on the Semantic Web, Synthesis Lectures on the Semantic Web: Theory and Technology, Morgan & Claypool, Palo Alto, CA, USA, 2012.

[7] S. Staab, R. Studer (Eds.), Handbook on Ontologies (2nd Ed.), Springer, 2009.

[8] P. Hitzler, M. Krötzsch, S. Rudolph, Foundations of Semantic Web technologies, Springer, 2010.

[9] L. Sinikallio, S. Drobac, M. Tamper, R. Leal, M. Koho, J. Tuominen, M. L. Mela, E. Hyvönen, Plenary debates of the Parliament of Finland as linked open data and in Parla-CLARIN markup, in: 3rd Conference on Language, Data and Knowledge, LDK 2021, Schloss Dagstuhl- Leibniz-Zentrum fur Informatik GmbH, Dagstuhl Publishing, 2021, pp. 1–17. URL: https://drops.dagstuhl.de/opus/volltexte/2021/14544/pdf/OASIcs-LDK-2021-8.pdf.

[10] P. Leskinen, E. Hyvönen, J. Tuominen, Members of Parliament in Finland knowledge graph and its linked open data service, in: of the 17th International Conference on Semantic Systems, 6-9 September 2021, Amsterdam, The Netherlands, 2021, pp. 255–269. URL: https://ebooks.iospress.nl/volumearticle/57420. doi:10.3233/SSW210049.

[11] K. Elo, J. Karimäki, Luonnonsuojelusta ilmastopolitiikkaan: Ympäristöpoliittisen käsitteistön muutos parlamenttipuheessa 1960–2020, Politiikka 63 (2021). doi:10.37452/politiikka.109690.

[12] H. Pokkimäki, P. Leskinen, M. Tamper, E. Hyvönen, Analyses of networks of politicians based on linked data: Case ParliamentSampo – Parliament of Finland on the Semantic Web, 2022. URL: http://seco.cs.aalto.fi/publications/2022/poikkimaki-et-al-2022.pdf, paper under peer review.

[13] E. Lapponi, M. G. Søyland, E. Velldal, S. Oepen, The Talk of Norway: a richly annotated corpus of the Norwegian parliament, 1998–2016, Lang Resources & Evaluation 52 (2018) 873–893. doi:10.1007/s10579-018-9411-5.

[14] A. Pancur, T. Erjavec, The siParl corpus of Slovene parliamentary proceedings, in: Proceedings of the Second ParlaCLARIN Workshop, European Language Resources Association, 2020, pp. 28–34. URL: https://www.aclweb.org/anthology/2020.509parlaclarin-1.6.

[15] M. La Mela, Tracing the emergence of Nordic allemansrätten through digitised parliamentary sources, in: M. Fridlund, M., Oiva, P. Paju (Eds.), Digital histories: Emergent approaches within the new digital history, Helsinki University Press, 2020, pp. 181–197. doi:10.33134/HUP-5-11.

[16] M. Lennes, FIN-CLARIN and language bank parliamentary data. Workshop 'Digital Parliamentary Data and Research', 2019. URL:

https://www2.helsinki.fi/en/helsinki-centre-for-digital-humanities/workshop-digital-parliamentary-data-and-research.

[17] A. Mansikkaniemi, P. Smit, M. Kurimo, Automatic construction of the Finnish parliament speech corpus, in: Proc. Interspeech 2017, 2017, pp. 3762–3766. doi:10.21437/Interspeech.2017-1115.

[18] M. Andrushchenko, K. Sandberg, R. Turunen, J. Marjanen, M. Hatavara, J. Kurunmäki, T. Nummenmaa, M. Hyvärinen, K. Teräs, J. Peltonen, J. Nummenmaa, Using parsed and annotated corpora to analyze parliamentarians' talk in Finland, Journal of the Association for Information Science and Technology 185 (2021) 1–15. doi:10.1002/asi.24500.

[19] C. Rauh, P. De Wilde, J. Schwalbach, The ParlSpeech data set: Annotated full-text vectors of 3.9 million plenary speeches in the key legislative chambers of seven European states (V1), 2017. doi:10.7910/DVN/E4RSP9.

[20] C. Reul, D. Christ, A. Hartelt, N. Balbach, M. Wehner, U. Springmann, C. Wick, C. Grundig, A. Büttner, F. Puppe, OCR4all – an open-source tool providing a (semi-) automatic OCR workflow for historical printings, arXiv preprint arXiv:1909.04032 (2019).

[21] K. Baierer, A. Büttner, E. Engl, L. Hinrichsen, C. Reul, OCR-D & OCR4all: Two complementary approaches for improved OCR of historical sources, in: 6th International Workshop on Computational History, CEUR Workshop Proceedings, 2021.

[22] S. Drobac, K. Lindén, Optical character recognition with neural networks and post-correction with finite state methods, International Journal on Document Analysis and Recognition (2020). doi:s10032-020-00359-9.

[23] K. T. Kettunen, J. M. O. Koistinen, et al., Open source Tesseract in Re-OCR of finnish fraktur from 19th and early 20th century newspapers and journals – collected notes on quality improvement, in: Digital Humanities in the Nordic Countries Proceedings of the Digital Humanities in the Nordic Countries 4th Conference, CEUR-WS.org, 2019.

[24] T. T. H. Nguyen, A. Jatowt, M. Coustaty, A. Doucet, Survey of post-ocr processing approaches, ACM Computing Surveys (CSUR) 54 (2021) 1–37.

[25] J. Z. Pan, Resource Description Framework, in: S. Staab, R. Studer (Eds.), Handbook on Ontologies, International Handbooks on Information Systems, Springer, Berlin, Heidelberg, 2009, pp. 71–90. doi:10.1007/978-3-540-92673-3_3.

[26] T. Erjavec, A. Pančur, Parla-CLARIN - a TEI schema for corpora of parliamentary proceedings, 2022. URL: https://clarin-eric.github.io/parla-clarin/.

[27] E. Mäkelä, LAS: an integrated language analysis tool for multiple languages., J. Open Source Software 1 (2016) 35. doi:10.21105/joss.00035.

[28] T. A. Pirinen, Omorfi—free and open source morphological lexical database for Finnish, in: Proceedings of the 20th Nordic conference of computational linguistics (NODALIDA 2015), 2015, pp. 313–315.

Results from rough data? The large-scale study of early modern historiography with multi-dimensional register analysis

Aatu Liimatta[1], Yann Ryan[1], Tanja Säily[1] and Mikko Tolonen[1]

[1]*University of Helsinki*

Abstract

Multi-dimensional register analysis is a methodology which can be used to extract functional dimensions from a set of texts. These dimensions describe various functional differences between the set of texts. The differences can be due to various situational constraints related to the production of the text, or they can be related to differences in the author's intent and communicative purpose. While this methodology has seen considerable use in contemporary linguistics, it has been less used in historical linguistics, and even less so in history, even though the ability to differentiate between various textual functions in historical data would be extremely useful and interesting from the point of view of a historian. In this paper, we perform a pilot study of multi-dimensional register analysis on a subset of texts from *Eighteenth Century Collections Online* (ECCO). In particular, our goal is to find out whether this kind of analysis is possible in the first place, or if it is hindered too much to be useful by the low quality of the ECCO data produced by optical character recognition (OCR). To do this, we first perform the analysis on ECCO data, after which we compare the results with results from running the same analysis on the same set of texts from ECCO-TCP, a manually cleaned subset of ECCO data. Our results show that not only are the results from the ECCO analysis interpretable, but they are also highly similar with the results from ECCO-TCP. Multi-dimensional register analysis appears to be a very promising and robust method which can work well even with low-quality data.

Keywords

register analysis, OCR issues, Scottish Enlightenment, historical writing, Eighteenth Century Collections Online (ECCO)

1. Introduction

Register analysis is a linguistic approach which examines language use in different situations and functions [1, pp. 6-7]. With methods such as the Multi-Dimensional Analysis, one can find dimensions of functional variation, which describe functional differences between texts in a dataset. However, studies of historical texts using these methods are rare. Such studies would provide valuable insight into historical linguistic variation.

DHNB2023 | Sustainability: Environment - Community - Data. The 7th Digital Humanities in the Nordic and Baltic Countries Conference. Oslo – Stavanger – Bergen, Norway. March 8–10, 2023.

aatu.liimatta@helsinki.fi (A. Liimatta); yann.ryan@helsinki.fi (Y. Ryan); tanja.saily@helsinki.fi (T. Säily); mikko.tolonen@helsinki.fi (M. Tolonen)

0000-0001-9056-1087 (A. Liimatta); 0000-0003-1878-4838 (Y. Ryan); 0000-0003-4407-8929 (T. Säily); 0000-0003-2892-8911 (M. Tolonen)

DHNB Publications, DHNB2023 Conference Proceedings, https://journals.uio.no/dhnbpub/issue/view/875

While historians have long relied on context to interpret meaning and placed a particular emphasis on linguistic acts, the full extent of the linguistic context in larger corpora has often been overlooked in the study of the process of history writing (historiography). Register analysis can provide valuable insights into how authors construct and convey their accounts of the past, but historians have not yet utilized this tool to its full potential. The purpose of this paper is to address these shortcomings and bridge these gaps.

Our article seeks to explore the potential of using register analysis for *Eighteenth Century Collections Online* (ECCO), a large but but noisy dataset of historical text, and to systematically examine the convergence of register analysis and the analysis of early modern historical writing. This intersection has the potential to produce far-reaching impacts, particularly when history and sociolinguistic analysis are brought together. By analyzing language use in different contexts, register analysis can shed light on how historical narratives were constructed and conveyed by authors of the past. Thus, our article aims to highlight the potential benefits of incorporating register analysis into historical analysis, emphasizing the importance of taking a comprehensive and interdisciplinary approach to the study of the past.

2. Background

Register analysis is a field of linguistics which focuses on the use of language in different situational contexts and for different purposes. Traditionally, register analysis focused on differences between *a priori* register distinctions, such as "formal" and "casual" or "written" and "spoken" registers. However, since the end of 1980s, so-called Multi-Dimensional Analysis (MDA), originally developed by Douglas Biber [2], has been extremely influential in quantitative corpus-based register analysis. MDA rejects the idea of such *a priori* register distinctions. Instead, through computational and statistical analysis of texts, it is possible to extract multiple dimensions along which various functional linguistic features vary in the dataset in question. By analysing the sets of features associated with the dimensions and the texts in which they appear, it is possible to find functional distinctions which differentiate the texts in the dataset.[1]

Register analysis applied to large historical datasets has great potential, offering the possibility of studying historical language use at a much greater scale than usually done. Historians have not frequently leveraged the benefits of extensive linguistic analysis to establish a broader framework for their research pursuits. The subject of this analysis, ECCO, serves as a good example. ECCO contains over 30 million pages of text, from over 200,000 documents, making it much larger than most hand-curated historical corpora. [3] In 2004 Gale claimed that ECCO contained every significant work in the English langugage printed in the eighteenth century, plus thousands of other important works [3, p. 56], though studies have shown that its representation of the entirety of the eighteenth century publishing landscape is uneven [4]. Nevertheless, a linguistic analysis which could confidently be applied to this dataset would provide new opportunities to study the wide variety of texts within ECCO, including pamphlets, legal

[1] Other methods can also be used for register analysis. Most register analysis methods are based on the idea of comparing the occurrences of functional linguistic features in different texts. We have chosen to use MDA in our analysis since it likely is the best-known and most widely used of these methods in linguistics. However, the findings will also be relevant for studies using other similar methods of register analysis.

documents and statutes, technical texts or instruction manuals, and non-elite writing. Register analysis, for instance, may allow us to understand more about the nature of writing about the past in the eighteenth century, by helping analyse changes in the style and communicative functions of historiography. History and other expository forms of writing moved towards professionalization throughout the century [5], and one hypothesis to be tested is that this may be reflected in the register dimensions.

However, while MDA has been used with great results for a long time (see [6][7]), only a handful of MDA studies have been conducted using historical data (e.g. [5][8][9]). Furthermore, these studies have used smaller, curated and hand-corrected corpora, meaning that the vast majority of historical text documents are excluded. ECCO and similar large-scale historical datasets are often not used in whole for linguistic or historical analyses, in part because of the low quality of the text data. ECCO text data has a large number of errors, for several reasons. First, the source texts themselves are of varying quality and suffer from printing issues such as bleed-through. Second, modern OCR engines are often not calibrated for the fonts used, and third, the OCR has been produced mostly from low-quality bitonal scans of microfilm. Hill and Hengchen [10] have shown that for a sample of ECCO, the mean token-level accuracy was 77%.

Our aim in this paper is to show that even in spite of these significant errors, certain types of of linguistic analysis, such as MDA, can produce robust results which compare surprisingly well to a much cleaner, transcribed set of texts. We do this by comparing the above-described dataset (ECCO-OCR hereafter) to another, the ECCO Text Creation Partnership (ECCO-TCP). ECCO-TCP[2], released in 2011, is a collection of approximately 2,100 documents, covering a range of books from the eighteenth century. [11] ECCO-TCP is double-keyed, meaning that each document was transcribed twice, and a third transcriber resolved disputes. While it does have errors, it is produced in a similar manner and likely has a similar error rate to most linguistic corpora. Originally conceived as part of a larger project [3] (a sister project, EEBO-TCP, contains over 40,000 documents), ECCO-TCP is nevertheless useful for comparing a clean dataset with the OCR version of the same dataset, because it has full overlap with texts from ECCO-OCR.

ECCO-TCP and OCR have been used for similar comparisons before. Hill and Hengchen [10] compare ECCO-OCR and ECCO-TCP with respect to a number of linguistic tasks, including topic modeling, authorship attribution, collocation analysis and vector space modeling. They find that topic modeling works quite well in ECCO-OCR, while collocation analysis can be more problematic unless great care is taken in the selection of subcorpora and research questions. Their study indicates that an OCR accuracy of 80% (in terms of F1 Scores calculated at the page level) is sufficient for most tasks, whereas a level below 70–75% significantly degrades the quality of the results. The greatest number OCR errors are found in words containing the long-s or ligatures. However, while Hill and Hengchen consider various bag-of-words approaches, we instead focus on differences in results provided by a multi-stage computational and statistical analysis procedure which relies on the correct identification of longer linguistic constructions.

[2] https://textcreationpartnership.org/tcp-texts/ecco-tcp-eighteenth-century-collections-online/

3. Data and methods

3.1. Datasets

As our source of data, we use the ECCO-TCP dataset and the equivalent set of texts from ECCO-OCR. For this proof of concept pilot analysis, we use the ECCO-OCR subset as our primary dataset, since in the present paper our goal is to test the ability to use the MDA methodology for analyses of large-scale historical text data despite the OCR quality issues. The ECCO-TCP dataset is then used as a point of comparison, to contrast the results of the ECCO-OCR analysis with a clean baseline to gauge the robustness of the methodology on less-than-perfect datasets.

In this study, we focus primarily on functional register differences between genres. To do this, we needed an existing, externally-produced set of genre labels which allowed us to analyse some kind of categorical differences between the texts. We used a set of genre labels generated using a state-of-the-art method [12]. This taxonomy of genre labels was custom-created to reflect sensible eighteenth-century generic distinctions, such as religious texts, histories, 'scientific' works, and so forth.

3.2. Data processing

First, both datasets, the ECCO-OCR subset and ECCO-TCP, were tagged for part of speech. The datasets were then analyzed for the linguistic features of interest. The algorithms to identify the features are based on Biber [2]. Most of Biber's original 67 features were included. However, a handful of the features are difficult to identify automatically, and so were not included in the analysis. Some of the features, viz. type-token ratio and mean word length, were excluded from the analysis because they are particularly vulnerable to low OCR quality. Furthermore, *first person pronouns* were split into singular and plural following some earlier studies (e.g. [13]), since the two features exhibit different functional behavior. In the end, the analysis included 58 features.

The number of occurrences of the features were then normalized. Normally, this would be done by dividing the number of occurrences of the feature by the number of running words or tokens in the text. However, because of the low quality of the OCR, the number of tokens extracted from the text by the part-of-speech tagging pipeline can be very far off from the "true" number. This is largely caused by the OCR process, which commonly splits individual words into multiple parts. Because of this, we instead normalized the feature counts by the number of *characters* in the text. The character count will also not be exactly correct, but it will be proportionally closer to the true value.

3.3. Multi-dimensional analysis

Since its inception, Multi-Dimensional Analysis (MDA) [2][14][1, p. 6] has been very influential in the field of quantitative corpus linguistics as a method for studying functional variation within language. The method is based on a simple idea: that linguistic features are functional, and therefore tend to be used more in the kinds of contexts for which they are particularly well-suited. Consequently, linguistic features well-suited for similar situational and functional contexts can be expected to appear in the same texts, when the situation or function of the text calls for them, and similarly be absent at the same time, when they are not needed.

By analyzing the co-occurrence patterns of a large number of linguistic features, it is possible to find groups of co-occurring features, which tend to appear in the same texts and be absent from the same texts. In practice, this is typically done using factor analysis, though other methods are also possible [15][16][17]. The groups of co-occurring features, i.e. the factors, can then be interpreted as functional dimensions of variation based on the assumption that the features in the group are present or absent due to some set of underlying communicative functions or situational concerns. The functional tendencies of e.g. textual genres or other groups of texts can then be compared in terms of their positioning on these functional dimensions.

In this study, the factor analysis was run twice. After the first run, any feature which did not have an absolute loading higher than .35 on any of the eight extracted factors was removed from the dataset, in order to reduce noise caused by features which are not central to the patterns of variation observed in the data. The factor analysis was then run again on the reduced feature set. The number of factors was decided based on an inspection of factor solutions with different numbers of factors, from four to nine. In the end, seven factors were extracted, as this solution was found to be the most readily interpretable.

After this, dimension scores were calculated for each text on each of the dimensions. First, the feature frequencies were standardized to a mean of 0 and standard deviation of 1, to ensure that high-frequency features would not drown out low-frequency features in the analysis. Then, as is commonly done in MDA studies to calculate the dimension score for a text on a dimension [2], the standardized frequencies in that text were added together for every feature which had a loading equal to or greater than 0.3 on that dimension.

4. Analysis

The dimension scores for the texts were then plotted across genres in ECCO to aid in the functional interpretation of the dimensions, and to enable inter-genre comparisons of functional tendencies. The positioning of the genres along the extracted dimensions is shown in Figure 1.

4.1. Dimension 1: Past/narrative/literary vs. non-past/speech-like focus

Table 1
Features associated with the positive and negative poles of Dimension 1

+	past tense, agentless passives, by-passives, perfect aspect, amplifiers, total prepositional phrases, downtoners, hedges, pied-piping relative clauses
-	contractions, present tense, discourse particles, other adverbial subordinators

Table 1 shows the features associated with Dimension 1. The main dichotomy between the positive and negative poles of Dimension 1 appears to be the distinction between *past tense* and *present tense*, indicating a functional division between a past temporal focus and a present or future temporal focus. This division is also supported by the presence of the *perfect aspect* on

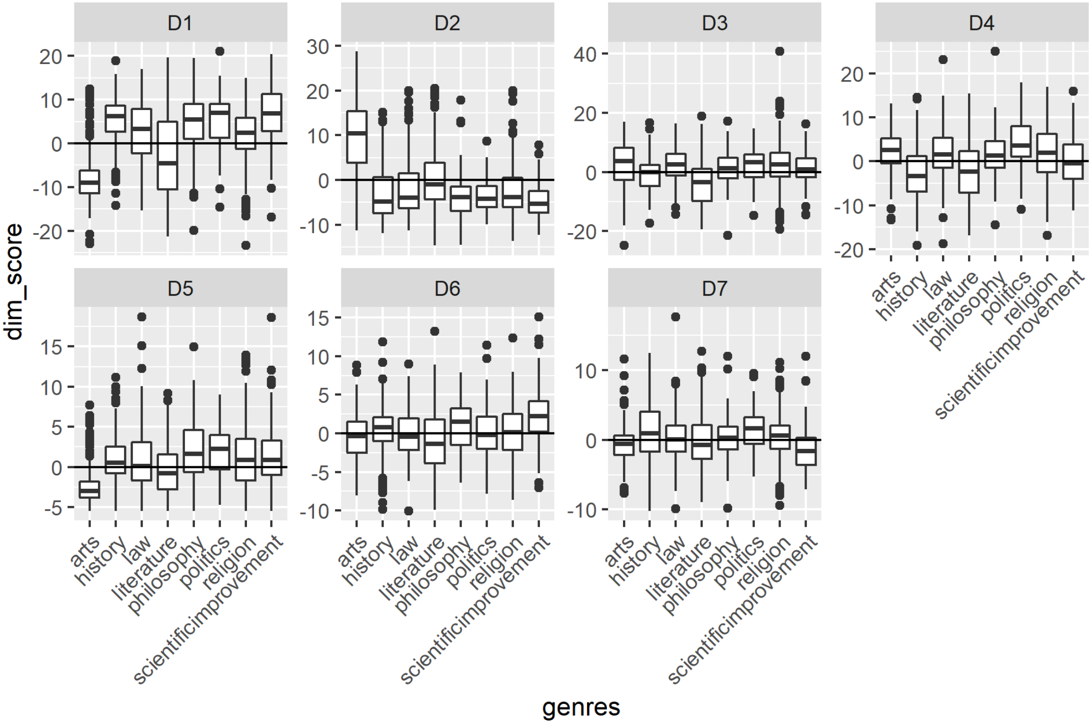


Figure 1: The positioning of the texts in the ECCO-OCR subset on the seven extracted dimensions by genre

the positive pole of the dimension. In this manner, Dimension 1 is somewhat aligned with the narrative universal register dimension proposed by Biber [6].

At the same time, many of the other features on the dimension point towards a different kind of distinction. Features such as *agentless passives*, *by-passives*, and *total prepositional phrases* are associated with a more abstract, impersonal, "literary" style, whereas the complementary features such as *contractions*, *discourse particles* and *other adverbial subordinators* are more typical of more "oral" or speech-like registers.

In Figure 1 we can see that the genre which differs the most from the rest on Dimension 1 is arts. The "arts" genre mostly contains works such as plays and poetry, which indeed tend to use the present tense and at the same time contain many more speech-like features than the average genre. Literature is more spread out on this dimension, because in addition to speech-like dialogue sections, it also contains narrative sections making use of the features of the narrative pole of the dimension, such as past tense. The other genres are more strongly associated with the narrative, literary side of the dimension.

Table 2
Features associated with the positive and negative poles of Dimension 2

+	first person singular pronouns, second person pronouns, subordinator-that deletion, private verbs, direct WH-questions, discourse particles, WH-clauses, perfect aspect, time adverbials, pro-verb do
-	total prepositional phrases, phrasal coordination

4.2. Dimension 2: Involved interpersonal focus

The features associated with Dimension 2 are shown in Table 2. The positive pole of the dimension is dominated by *first person singular pronouns* and *second person pronouns*, which point towards an involved, interpersonal function for this dimension. These functions are also supported by many of the other features on the dimension. For instance, *private verbs*, i.e. verbs which express internal mental states, such as *think*, *feel*, and *believe*, are often used in contexts where the speaker or writer is involved in the produced text. Similarly, *direct WH-questions* imply an interactive or interpersonal focus. Features such as *WH-clauses* and *time adverbials* help locate and refer to a specific referent or time.

On Biber's well-known first dimension [2], this kind of involved register is placed in contrast with informationally dense registers. Dimension 2 in the present study also bears some signs of such a dichotomy. Both of the features on the negative pole of the dimension can be considered to contain a higher informational load; on the positive pole, instead, *pro-verb "do"* replaces a whole verb phrase, reducing the informational load of the text.

In Figure 1, arts again have a clearly different score on Dimension 2 compared to the other genres. Plays contain a high number of spoken lines, which tend to be very involved and interpersonal in nature, and poetry is also similar in many ways. The other genres are very different along this dimension, even literature, which is still the closest to arts with its higher number of characters' lines but a large proportion of non-spoken description when compared to arts.

4.3. Dimension 3: Static statement

Table 3
Features associated with the positive and negative poles of Dimension 3

+	be as main verb, existential there, predicative adjectives, pronoun it, pro-verb do, indefinite pronouns, demonstrative pronouns, causative adverbial subordinators: because, conditional adverbial subordinators: if & unless
-	attributive adjectives, other adverbial subordinators

Table 3 lists the features associated with Dimension 3. The main feature on this dimension is *"be" as main verb*. The verb *be* is also closely linked with other features on the dimension. For example, *existential "there"*, i.e. the construction "there is/there are", is formed using the verb *be*, and similarly *predicative adjectives* are adjectives occurring in a predicative position, i.e. following the verb *be*, such as in the sentence “The house is big” (as opposed to *attributive adjectives*, such as in the phrase “the big house”).

In general, the positive pole of this dimension appears to describe texts which tend to use more predicative, or static, expressions. In other words, texts in the positive pole of the dimension express more than average the nature of something, that *something is (like) something*, or that *there is something somewhere*, as opposed to using other verbs to express what that something does, or what is done to it.

While the placement of the genres on this dimension is relatively level in Figure 1, there are still small differences in their tendencies. Arts, law, philosophy, politics and religion as well as scientific improvement appear to use this kind of description slightly more, whereas history and literature use it slightly less. This can be explained as history and literature having slightly more focus on active actions, whereas the other genres are slightly more interested in the nature of things. However, as can be seen in the figure, this difference is quite subtle.

4.4. Dimension 4: Expression of options and possibilities

Table 4
Features associated with Dimension 4

+	possibility modals, necessity modals, predictive modals, split auxiliaries, present tense, analytic negation: not, conditional adverbial subordinators: if & unless, infinitives, suasive verbs

The main features associated with Dimension 4, listed in Table 4, are clearly the three types of modal verbs: *possibility modals* (such as *can* or *may*), *necessity modals* (such as *must* or *should*), and *predictive modals* (such as *will* or *would*). These features already clearly position texts scoring highly on this dimension as talking about various options, necessities and possibilities. The other features on the dimension support this function. For instance, *analytic negation “not”* is used to reverse the modality of the three modal verb classes to express that something can, must, or will *not* be done. Similarly, *conditional adverbial subordinators “if” & “unless”* are used to qualify and limit the scope of the modal content.

This dimension also contains all of the features included on Biber’s original Dimension 4 [2], which he labeled *Overt expression of persuasion* or *Overt expression of argumentation* [6]. However, Dimension 4 of the present study also includes features not included on Biber’s original dimension. Still, it is true that the dimension in this study, too, has a certain connection to persuasion and argumentation. We have tentatively given the dimension a wider label, *Expression of options and possibilities*.

In Figure 1, genres such as politics and religion are naturally placed high on this dimension, and even law, expressing the various consequences and means of dealing with different legal situations. On the other hand, history and literature have clearly lower scores on this dimension: they express how things were or what happened in the past or in the fictional world of a story, and so these genres do not need so many hypotheticals.

Table 5
Features associated with Dimension 5

+	WH relative clauses on object positions, WH relative clauses on subject position, that adjective complement, that relative clauses on object position, pied-piping relative clauses

4.5. Dimension 5: Complex reference

The features associated with Dimension 5, shown in Table 5, are all complex clauses acting as subjects, objects, or modifiers. These features, such as *WH relative clauses on object positions* (e.g. "the question which was the main topic of discussion yesterday"), *pied-piping relative clauses* (e.g. "the manner in which it was done"), and *"that" adjective complements* (e.g. "I am glad that you think so"), all help to integrate large amounts of information and expressions of complex relationships between ideas, actors and objects into more condensed textual units.

In Figure 1, genres such as philosophy, politics, religion and scientific improvement are positioned relatively high on this dimension. These genres deal with topics which are often quite complex, and as such can make use of constructions which refer to complex referents. Other genres, such as history and law, have slightly lower scores, but also potentially a great deal of internal variation. Finally, literature and particularly arts do not typically use constructions like these that much.

4.6. Dimension 6: Evaluation and qualification of information

Table 6
Features associated with Dimension 6

+	total adverbs, emphatics, hedges, amplifiers, downtoners

All of the features on Dimension 6, listed in Table 6, are words which in some way evaluate or qualify information. *Emphatics*, *hedges*, *amplifiers*, and *downtoners* are used to either intensify or reduce the certainty or strength of the claim being made. Similarly, *adverbs* express the manner of performing an action, and in that way qualify or evaluate the content of the expression.

Figure 1 shows that most genres do not exhibit very large differences in terms of Dimension 6. However, scientific improvement and philosophy are on average slightly above the others. After all, evaluating and qualifying various ideas and claims is a central part of both fields.

4.7. Dimension 7: Third-person focus

Table 7
Features associated with Dimension 7

+	third person personal pronouns, public verbs, infinitives, suasive verbs, perfect aspect

Dimension 7 features, as shown in Table 7, are all associated with the focus being on a third-person actor. While *third person personal pronouns* are the clearest sign of this, other features on this dimension also support the function of third-person focus. Most importantly, public verbs (i.e. verbs which express publicly observable actions, such as *say* or *explain*) and suasive verbs (i.e. verbs which imply intention to bring about change, such as *command* or *recommend*), clearly focus on the actions of a third-person actor.

Most genres do not differ that much on Dimension 7 in Figure 1. However, a couple of the genres have slight differences compared to the others. Most importantly, scientific improvement is below the others, focusing more on science and engineering than any person. On the other hand, both history and politics are placed slightly higher than the rest, since these genres often focus on historical characters or political actors and their deeds.

5. OCR vs. TCP

As shown in the previous section, the features associated with the seven dimensions form largely meaningful groups whose function and use is quite readily interpretable based on this analysis. This result is already very promising in terms of the robustness and usability of the MDA methodology even on less-than-perfect OCR data, and indicates that at least some useful results can be acquired from imperfect datasets.

However, based on these results only, it is unclear how similar the dimensions extracted from the imperfect OCR dataset are to dimensions which would be extracted from a clean version of the same dataset. In order to gauge the similarity of this result using the imperfect dataset to a result from a clean dataset, we performed the exact same analysis steps, including part-of-speech tagging, feature identification, and MDA with the same parameters, on the ECCO-TCP set of texts. Since the original analysis was based on the subset of texts from ECCO-OCR which matches the texts in ECCO-TCP, we can get a good idea of what the results would look like if the exact same analysis process was performed on a clean version of the same dataset.

The results from the analysis of the TCP dataset are very promising from the outset, when compared to the results of the OCR subset analysis. The dimensions extracted from the TCP dataset are in a slightly different order, reflecting small changes in their relative importance, but the dimensions comprise largely the same features as the OCR dimensions, making it trivial to map the TCP dimensions to the OCR dimensions one-to-one.[3] Importantly, while there are small number of added or removed features, the features most strongly associated with the dimensions do not change; all changes in features on the dimensions take place among the features with weaker loadings on the corresponding factors.

But how much is the positioning of the texts and genres on the dimensions affected by these small differences in the feature makeup of the dimensions, the differences in the part-of-speech tagging, and the corresponding differences in the identification of the features in the texts? Figure 2 plots the results of the TCP analysis alongside the results of the OCR analysis in order to see how much the positioning of the genres differs on the seven dimensions. Based on this figure, it seems clear that the patterns are quite similar between the two datasets. There are small differences, but the dimensions clearly capture similar phenomena in the two datasets.

[3] The TCP dimensions have been labeled here with the numbers of the equivalent OCR dimensions to simplify the comparison of the dimensions extracted from the two datasets.

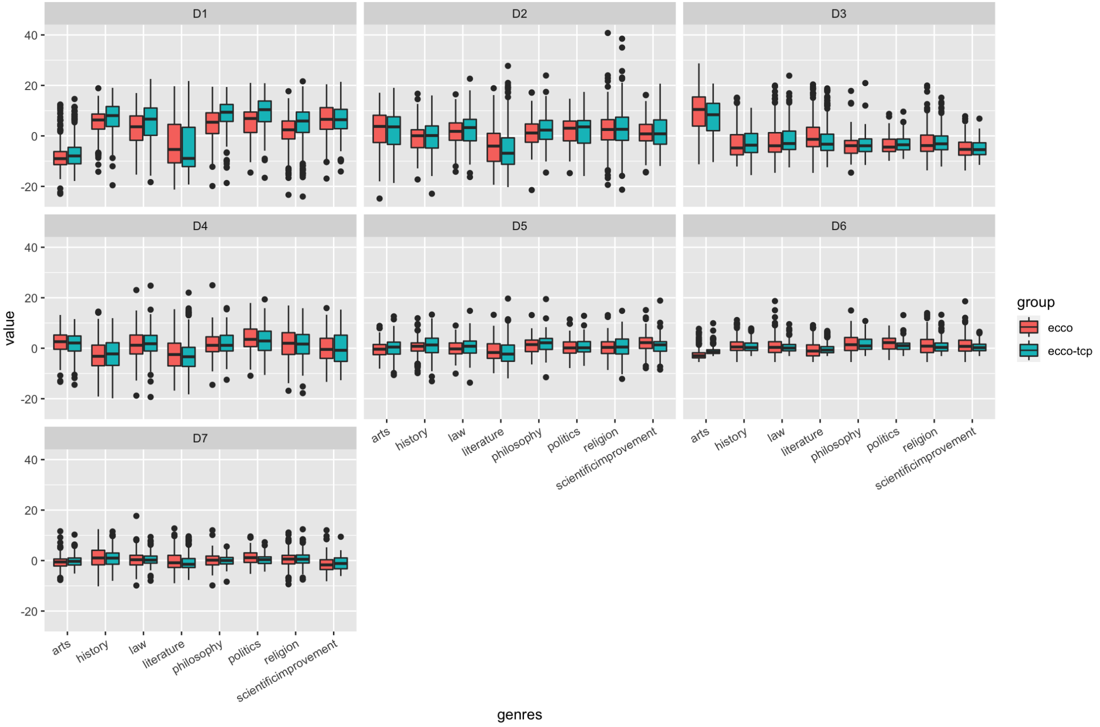


Figure 2: A comparison of the positioning of the texts in two MDA analyses, one based on ECCO-TCP (blue) and one on the equivalent ECCO-OCR subset (red). Every facet shows the comparison for one dimension. Each pair of boxes shows the distributions of the dimension scores based on the two analyses. A similar vertical positioning of a pair of boxes means that the analysis produces similar results for the two datasets on that dimension for that genre.

To further illustrate the similarity of the dimensions extracted from the OCR and TCP datasets, Table 8 shows the strong correlation of the dimension scores from the OCR and TCP analyses for all seven dimensions in all eight genres. That is, for any of the dimensions, a text scoring highly on that dimension in the OCR dataset is also very likely to score highly on the equivalent dimension in the TCP dataset, and vice versa.

MDA requires automated identification of complex linguistic features, and it would stand to reason that the low OCR quality of ECCO-OCR would be extremely detrimental to such analysis. However, it is not entirely unexpected that the method works even with this kind of data. A wide variety of MDA studies have used different sets of features. While the dimensions they have produced are not the exact same, the dimensions are nevertheless typically "compatible" with each other [18], meaning that they capture similar kinds of linguistic phenomena and functions even if their precise structure differs. For our puroses, if different sets of features produce similar results, it means that some results can still be expected from the analysis even if not all features in the set of features are identified correctly, such as because of OCR errors. MDA also mitigates the detrimental effect of OCR errors in another way: because the method works based on the correlations of the presence and absence of the features. Even if every

Table 8
P-values for the correlation between the dimension scores for ECCO OCR and TCP by genre

Genre	D1	D2	D3	D4	D5	D6	D7
Arts	0.923	0.958	0.941	0.883	0.877	0.739	0.686
History	0.702	0.887	0.948	0.932	0.767	0.763	0.898
Law	0.932	0.932	0.925	0.892	0.876	0.852	0.732
Literature	0.951	0.945	0.853	0.900	0.894	0.844	0.884
Philosophy	0.899	0.937	0.886	0.855	0.680	0.931	0.718
Politics	0.926	0.911	0.808	0.907	0.900	0.811	0.657
Religion	0.930	0.948	0.862	0.929	0.899	0.857	0.754
Scientific improvement	0.935	0.935	0.771	0.930	0.882	0.878	0.817

feature is only identified a small fraction of the times it actually appears, its correlation patterns should not change much: when the feature is identified correctly, it still appears in the same texts with other features which share its communicative functions.

6. Discussion

Historians have traditionally relied on context as their primary tool, but linguistic context has often been overlooked in their analytical toolkit. Register analysis involves examining language use in different contexts and identifying the features that characterize them. Despite its potential to provide valuable insights into the historical events and texts under examination, register analysis has rarely been utilized to its full potential by historians.

As our analysis shows, even when working with imperfect historical datasets, the multi-dimensional method of register analysis (MDA) produces dimensions “compatible” [18] with e.g. the register universals [6]. The method can therefore also be useful for providing insight into functional variation in historical texts for historians and linguists alike. By analyzing texts in terms of register dimensions and the features associated with them, we can find out much more about their various functions. For instance, when applied to historical narratives, register analysis can shed light on the ways in which authors use language to construct and convey their accounts of the past. In higher-level analyses, we can consider the dimensions and their uses in historical writing, and see to which degree the various functions appear in the texts we are interested in. But it is also possible to zoom in from this high-level picture, and analyze in detail the occurrences of the features associated with those dimensions in the texts.

For example, by analyzing texts which place high on Dimension 1, the past/narrative/literary vs. non-past/speech-like dimension, in terms of how they use the features associated with the dimension, such as past tense, agentless passives, and by-passives, we can gain insight into how authors attribute actions and events to particular actors or causes. In a similar manner, the other features on the dimension, as well as the other dimensions and their features, can also be indicative of an author’s stance towards the events they describe. Still other features, such as those associated with Dimension 5, can affect the flow and coherence of the narrative. By examining all of these different features and their distributions across historical texts, register analysis can help us better understand the ways in which language is used to construct historical narratives. Many of the features of dimensions 1 and 5 are demonstrated (highlighted with bolding and italics, respectively) in the following passage from a text in the "history" genre:[4]

> He **summoned** a Parliament, ***to whom*** he **made** bitter complaints **against** the irruption of the Scotch, the absurd imposture *which* **was countenanced by** that nation, the cruel devastation *which* they **had spread over** the northern counties, and the complicated affront *which* **had thus been offered** both **to** the King and kingdom of England. (David Hume, 1759, *The history of England*)

Another important aspect of register analysis in historical narratives is the use of Dimension 3, the static statement dimension. This involves examining the ways in which language is used to make factual claims and convey information. For example, texts scoring high on this dimension may have a more declarative or authoritative tone, arising from simple statements of fact, e.g. using the verb *be* as a main verb as well as existential *there* and predicative adjectives, and the use of the pronoun *it* to emphasize the importance or significance of certain events or objects. Pro-verb *do*, indefinite and demonstrative pronouns, as well as causative and conditional adverbial subordinators, can all contribute to the overall clarity and specificity of the statement. Additionally, the use of other adverbial subordinators and attributive adjectives can further refine or contextualize the information being presented. By analyzing these linguistic features in the context of historical narratives, register analysis can help identify the ways in which authors use language to make claims and convey information in a persuasive and authoritative manner. Many of the features associated with Dimension 3 can be seen in the following two examples, from the "history" and "religion" genres respectively, using these features to build an authoritative tone and describing the nature of things.

> **There is** no possible cafe, either of immorality or even inconvenience, but what **is** within the reach and correction of the COMMON LAW; for, **it is** a rule therein, that " **nothing** which **is** against REASON **is lavfull**;- " and, fureiy, every thing that **is immoral is** " againf reason ;" and again, by another rule, " **nothing** (that **is inconvenient is lawful**.§" (Granville Sharp, 1784, *An account of the ancient division of the English nation into hundreds and tithings*)

> With refpet to what are called denominations of religion, **if** every one is left to judge of **its** own religion, **there is** no such thing as a religion that **is wrong**; but **if** they are to judge of each others religion, **there is** no such thing as a religion that **is right**; and therefore, all the world **is right**, or all the world **is wrong**. (Thomas Paine, 1791, *Rights of man*)

A crucial application of register analysis for historians would be the development of a large-scale automated system that could tag longer narrative sections in multiple texts on a similar topic and automatically detect instances where the register shifts to a more declarative or statement-like register, indicating a distinct type of activity within the framework of historical

[4] All examples in this paper have been copied directly from the OCR text files and include the original OCR errors.

writing. This would be an invaluable tool for historians seeking to analyze differences among early modern histories that follow common practices, enabling them to pinpoint those historians who break the narrative mold when discussing common events or themes. By taking the historian's analysis to a new level, such an automated system could help uncover hidden insights and provide a more nuanced understanding of what particular authors were doing when writing about historical events and narratives.

Furthermore, the use of static statement in historical narratives can reveal a distinct aspect of history known as 'universalism', the tendency of authors to use history as a means of uncovering universal truths about human nature and progress, rather than simply recounting past events, for example the teleological approach taken by many eighteenth-century historians [19]. For instance, an author who makes a statement rather than providing a detailed explanation of a particular event or phenomenon may be attempting to convey a broader, more abstract idea. By using language to make categorical claims and assert general principles, authors can create a sense of universality that transcends individual historical contexts. Register analysis can give us insight into the ways in which authors use language to construct and convey these broader historical narratives, and the implications that this has for our understanding of the past.

Our next objective is to embark on large-scale analysis of historiography, starting by compiling a comprehensive and uniform corpus of relevant texts. Developing a method for tagging registers of interest, such as narrative and statement-like registers, will enable us to analyze the corpus at a larger scale. Our analysis will then focus on particular cases, allowing us to conduct a more nuanced examination of how specific early modern historians use registers. This endeavor will help advance the state of the art in the field and generate new knowledge about crucial differences between authors writing about British history during the Stuart era and beyond.

7. Conclusion

Our paper demonstrates that even when working with datasets suffering from significant problems with low-quality OCR, the multi-dimensional method of register analysis (MDA) can still be used to good effect. Through analysis of the extracted functional dimensions, we have shown that these dimensions are not only meaningful and interpretable in themselves, but also within the context of the genres present in the dataset. This is of particular relevance to the analysis of early modern historiography, which was a focus of our study. By highlighting the potential benefits of MDA in this context, our paper contributes to a growing body of literature demonstrating the value of interdisciplinary approaches to historical analysis. Furthermore, the comparison of these dimensions with the equivalent dimensions from the clean ECCO-TCP dataset shows that there are only minor differences between the dimensions extracted from the clean and dirty datasets and the positioning of the texts along those dimensions, particularly when the results are analyzed in aggregate. Overall, the two analyses capture similar phenomena, strongly supporting our hypothesis that MDA is a robust methodology which works well even with lower-quality data. Due to the many issues with the underlying data, the OCR quality of large datasets is unlikely to substantially improve in the near future. Consequently, any methods which can work reasonably well with such low-quality data are desirable.

The ultimate goal of our work with register analysis is to enable its application to full ECCO datasets, and to utilize this approach in the analysis of early modern historiography from a historical perspective – an area which has yet to be fully explored. By demonstrating the value of this approach, we hope to encourage more researchers to incorporate register analysis into their historical analysis and contribute to a more nuanced understanding of the past.

Acknowledgments

This study is a part of the Academy of Finland funded project *Rise of Commercial Society and Eighteenth-century Publishing* (RiCEP; grant numbers 333716 and 333717).

References

[1] D. Biber, S. Conrad, Register, genre, and style, Cambridge University Press, Cambridge, 2009. doi:10.1017/CBO9780511814358.

[2] D. Biber, Variation across speech and writing, Cambridge University Press, Cambridge, 1988. doi:10.1017/CBO9780511621024.

[3] S. H. Gregg, Old books and digital publishing: Eighteenth-Century Collections Online, 1 ed., Cambridge University Press, 2021. doi:10.1017/9781108767415.

[4] M. Tolonen, E. Mäkelä, L. Lahti, The anatomy of Eighteenth Century Collections Online (ECCO), Eighteenth-Century Studies 56 (2022) 95–123. doi:10.1353/ecs.2022.0060.

[5] D. Biber, E. Finegan, Diachronic relations among speech-based and written registers in English, in: T. Nevalainen, L. Kahlas-Tarkka (Eds.), To explain the present: Studies in the changing English language in honour of Matti Rissanen, Société Néophilologique, Helsinki, 1997, pp. 66–83.

[6] D. Biber, Using multi-dimensional analysis to explore cross-linguistic universals of register variation, Languages in Contrast 14 (2014) 7–34. doi:10.1075/lic.14.1.02bib.

[7] S. Conrad, D. Biber, Variation in English: Multi-dimensional studies, Pearson Education, Harlow, England, 2001. doi:10.4324/9781315840888.

[8] D. Biber, J. Burges, Historical change in the language use of women and men: Gender differences in dramatic dialogue, Journal of English Linguistics 28 (2000) 21–37. doi:10.1177/00754240022004857.

[9] D. Biber, Dimensions of variation among eighteenth-century speech-based and written registers, in: S. Conrad, D. Biber (Eds.), Variation in English: Multi-dimensional studies, Pearson Education, Harlow, England, 2001, pp. 200–214.

[10] M. J. Hill, S. Hengchen, Quantifying the impact of dirty OCR on historical text analysis: Eighteenth Century Collections Online as a case study, Digital Scholarship in the Humanities 34 (2019) 825–843. doi:10.1093/llc/fqz024.

[11] The results of keying instead of OCR – Text Creation Partnership, n.d. URL: https://textcreationpartnership.org/using-tcp-content/results-of-keying/.

[12] J. Zhang, Y. C. Ryan, I. Rastas, F. Ginter, M. Tolonen, R. Babbar, Detecting sequential genre change in eighteenth-century texts, in: F. Karsdorp, A. Lassche, K. Nielbo (Eds.), Proceedings of the Computational Humanities Research Conference 2022, volume 3290

of *CEUR Workshop Proceedings*, CEUR, Antwerp, Belgium, 2022, pp. 243–255. URL: https://ceur-ws.org/Vol-3290/#short_paper2630.

[13] A. Liimatta, Register variation across text lengths: Evidence from social media, International Journal of Corpus Linguistics (2022). doi:10.1075/ijcl.20177.lii.

[14] D. Biber, S. Conrad, Introduction: Multi-dimensional analysis and the study of register variation, in: S. Conrad, D. Biber (Eds.), Variation in English: Multi-dimensional studies, Pearson Education, Harlow, England, 2001, pp. 3–12.

[15] I. Clarke, J. Grieve, Dimensions of abusive language on Twitter, in: Z. Waseem, W. Hui Kyong, D. Hovy, J. Tetreault (Eds.), Proceedings of the First Workshop on Abusive Language Online, Association for Computational Linguistics, Vancouver, BC, 2017, pp. 1–10. doi:10.18653/v1/W17-3001.

[16] J. Egbert, D. Biber, Do all roads lead to Rome?: Modeling register variation with factor analysis and discriminant analysis, Corpus Linguistics and Linguistic Theory 14 (2018) 233–274. doi:10.1515/cllt-2016-0016.

[17] I. Clarke, J. Grieve, Stylistic variation on the Donald Trump Twitter account: A linguistic analysis of tweets posted between 2009 and 2018, PLoS ONE 14 (2019). doi:10.1371/journal.pone.0222062.

[18] T. McEnery, A. Hardie, Corpus linguistics: Method, theory and practice, Cambridge University Press, Cambridge, New York, 2012. doi:10.1017/CBO9780511981395.

[19] M. G. H. Pittock, Historiography, in: A. Broadie, C. Smith (Eds.), The Cambridge companion to the Scottish Enlightenment, 2 ed., Cambridge University Press, 2019, pp. 248–270. doi:10.1017/9781108355063.015.

Exploring the Sentiment of Latvian Twitter Food Posts in Various Weather Conditions

Maija Kāle[1], Matīss Rikters[2]

[1]Faculty of Computing, University of Latvia
[2]National Institute of Advanced Industrial Science and Technology

Abstract

Food choice is a complex phenomenon influenced by factors such as taste, environment, culture, weather and many others. Although people spend most of their lives indoors, weather conditions remain influential, both in shaping seasonal food cultures in particular geographical areas and in influencing individual choices. With the recent increase in the availability of datasets on food and its perception as reflected in Twitter and historical weather data, we seek to explore food-related tweets in different weather conditions. In this paper, we examine a Latvian food tweet dataset covering the last decade in conjunction with a weather observation dataset consisting of average temperature, precipitation and other phenomena. We find out which weather conditions lead to specific food information sharing; we automatically classify tweet sentiment and discuss how it changes depending on the weather. We also explore the dynamics of sentiment related to meat and meat consumption on Twitter over a ten-year period. The rationale for focusing on tweeters' sentiments about different meat-containing foods is due to the emergence of new discourses related to food consumption - the meat industry's impact on planetary health, the levels of biodiversity, pollution and CO2 that influence and shape climate change, as well as the planet's ecosystems as a whole.

Keywords

Linguistics, Social Network, Analysis, Food

1. Introduction

Food choice and consumption play an important role in public health. Obesity, type 2 diabetes and cardiovascular diseases are just some of the health problems associated with poor diet. According to the WHO Global Health Observatory (2016), one in four adults is overweight and one in ten is obese. The global prevalence of obesity has reached pandemic level. Therefore, it is of utmost importance to understand the underlying factors of food choice, which is a complex process influenced by various endogenous factors such as taste, quality, texture, colour and others, as well as exogenous or external factors ranging from demography, education level, time of day, weather, the environment in which food is consumed and others [1, 2, 3, 4].

Although most of our modern lives are spent indoors, weather and climate conditions still influence our food preferences and consumption [5]. Sunny weather and moderate temperatures

DHNB2023 | Sustainability: Environment - Community - Data. The 7th Digital Humanities in the Nordic and Baltic Countries Conference. Oslo – Stavanger – Bergen, Norway. March 8–10, 2023.

name.surname@lu.lv (M. Kāle); name.surname@aist.go.jp (M. Rikters)
0000-0002-6951-9009 (M. Kāle); 0000-0002-3530-6873 (M. Rikters)

DHNB Publications, DHNB2023 Conference Proceedings, https://journals.uio.no/dhnbpub/issue/view/875

lead to better moods among food consumers, while more extreme weather (hot, cold, any precipitation) means less pleasant weather conditions that affect mood and thus food consumption experiences. This is important in understanding that mood is the determining factor in food choices, with good mood associated with healthier food choices and bad mood associated with less healthy food choices [3].

While the impact of food on personal health is an area discussed by food policy makers and nutritionists globally, another new discourse has emerged in relation to food consumption, namely, the impact on planetary health or levels of biodiversity, pollution and CO2 that influence and shape climate change, as well as the planet's ecosystems overall [6]. One third of global carbon dioxide emissions are assigned to food systems, where the largest contribution comes from agriculture and land-use activities (estimated 71% of the total emissions), while the food supply chain - transport, consumption, retail and other related processes account for 29% respectively [7]. Meat production makes the largest impact when it comes to producing greenhouse gases, as it accounts for nearly 60% of all greenhouse gases from food production [8]. Beef accounts for one quarter of the total emissions, and in general, the use of animals for meat causes twice the pollution of producing plant-based foods [8].

The high emissions caused by meat production in the context of climate change have meant that the current food start-up and innovation scene is dominated by ideas focused on developing alternative plant-based proteins. The discourse on the future of food is therefore largely about the future of meat. Alternative proteins, lab-grown meat, vegan diets and flexitarian lifestyles - all of these concepts contribute to the discussion of how or whether we will consume meat in the future. Despite a fairly unified political push towards reduced meat consumption [9], the issue is becoming increasingly polarised at the level of social sentiment [6]. While there is increasing investment in meat replacement innovation, there is currently no evidence of a mass shift away from meat consumption globally.

Given the complexity of food consumption and our willingness to illustrate this complexity, we choose to focus our analysis on weather data and meat consumption. We justify this choice by illustrating that both areas are under-researched and should be considered holistically, taking into account the complex nature of food choices as such. We also use an under-utilised resource in food analysis - big data from social media, particularly Twitter. Social media in general is one of the best places to track sentiment around specific food consumption, where food is widely documented and discussed in multiple formats [10]. The analysis of social network data has become popular in consumer studies where language data is analysed. Food emerges as one of the key topics discussed on Twitter, the social network we focus on in our work. As a platform primarily for text rather than images, and because of its accessibility for research purposes, Twitter is the digital space where it has become possible to track the most random details of tweeters' everyday lives - including information about what, how and where they eat [11].

The driving motivation for this research is to build a better understanding of the world, in particular by looking at food consumption and the exchange of food-related information on social media. Food choices made by consumers have a major impact on public health and the sustainability of the planet, but due to the interdisciplinary nature of food, many important issues have been under-researched in narrowly focused research disciplines. This research aims to fill this gap and provide a methodology focused on sentiment analysis to understand food consumers, given the role that social media plays in modern lifestyles. With our approach, we aim to contribute to a growing area of research that focuses on interdisciplinary research questions and insights into the future of food [12]. The collection of food-related data is an unmet challenge, so innovative ways of using social media and other large-scale data are the key innovative approach that this research offers.

2. Research Focus

We chose the social network Twitter for the analysis due to the availability of a large food-related data corpus in Latvian language that has been recently published - the Latvian Twitter Eater Corpus (LTEC [13]). We focused on a specific social media - Twitter, where food is one of the main topics discussed, providing us with spontaneous reactions of food consumers, which is a unique feature compared to other data collection methods such as reviews or food diaries [14]. Our analysis of the LTEC provided a series of food-related discussions that we could correlate with the weather data, leading to the following research questions 1) Is there a correlation between food tweet sentiment and the weather the tweet authors are experiencing at the time of tweeting? 2) Are there differences in the frequency of food mentioned in tweets depending on the weather, and if so, what are the differences? As Twitter is a digital space where food experiences are shared instantly, we can better explain the context in which tweet authors share their thoughts with our analysis of weather data. Given previous studies that have demonstrated the link between weather, mood and food perception, our work aims to illustrate this link through tweet sentiment analysis. We refine our study by looking at frequencies - which food authors tweet more in pleasant weather and unpleasant weather conditions, mapping the weather-related food scene in Latvian language Twitter. With this analysis of weather-related dynamics in LTEC, we contribute to the field of research on the impact of weather on food consumption for the geographical region of Latvia and contribute to a broader understanding of the impact of weather on food consumers globally.

We focused our analysis on all entries related to meat and meat products found in food-related tweets. The sentiment of the tweets was then analysed in terms of their positive, neutral or negative valence. Bearing in mind that social media is generally considered to present mostly positive experiences [15], we carried out a general sentiment analysis of the food tweets. We looked at the representation of meat and meat products in time dynamics, trying to capture both the historical and contemporary ’zeitgeist’ in relation to meat consumption as it is represented in the Latvian-speaking community. In addition to an analysis of meat, we also looked at the representation of vegan and vegetarian food on Twitter, as well as debates about alternative proteins.

When analysing the discourse on meat in social media, it is important to be aware of the context of the society in question. When it comes to Latvian food culture and national cuisine, there is general agreement that the basis of Latvian cuisine is potatoes, dairy products, fish and meat, especially pork. It has developed as a heritage of peasant food in a mixture with aristocratic influence, similar to other European countries. [1] A slightly better understanding of food consumption can be based on the different seasons that Latvian society goes through during

[1] News portal of Latvian Radio and Television
https://eng.lsm.lv/article/culture/food-drink/traditional-and-national-latvian-foods-whats-the-difference.a466155/

the year. Latvia also has its seasonal food preferences, as depicted on social media: grey peas, tangerines and gingerbread during the Christmas season, and cold soup, strawberries and ice cream during the hot spring and summer [11].

When analysing the discourse of meat in social media, it is important to be aware of the context the particular society is operating in. When it comes to Latvian food culture and national cuisine general agreement is that the basis of Latvian cuisine is potato, dairy products, fish and meat, and pork in particular. It has formed as a heritage of peasant food in melange with aristocratic influence, similarly as in other European countries [2] Somewhat better understanding of food consumption can be based on different seasons that Latvian society lives through during the year. Latvia also has its seasonal food preferences as depicted in social media: grey peas, tangerines and gingerbread during Christmas time, and cold soup strawberries and ice cream during hot spring and summer time [11].

As a northern European country, Latvia has four distinct seasons where autumn and winter are relatively cold, dark and rainy, while summers are short and warm. What regards food choices, any society is sensitive to temperature and weather fluctuations, which is particularly evident in countries with greater seasonal variations in temperature [3]. Thus, Latvia is an example with various weather conditions that can be analysed from the perspective of tweeting about food: winter lasts from December until February, spring from March until May, summer is from June until August and finally, autumn from September until November. The average annual air temperature in Latvia is only +5.9°C. The year's warmest month is July, and the coldest months are January and February. February is also the snowiest month of the year there. The months with the most precipitation are July and August, while the least is in February and March. The highest wind speeds are in November, December and January, and the lowest wind speeds are in July and August. The months from May to August have the most days of sunshine, while in November, December and January, the Sun shines on average only 2-3 hours a day [16].

3. Related Work

In this section, we will review research that links weather and food data, as well as research related to meat consumption and perception. We will first look at studies of weather-related data, which are few and far between, indicating the difficulty of using big data in food-related research [17].

Weather people is a term used by Bakhshi [18] to explain our dependence on the weather for food choice and satisfaction. While the weather is known to significantly alter consumers' moods and consequently their behaviour [4], there have been surprisingly few studies illustrating the weather's impact on food perception and choice, except for some that have used online and offline restaurant reviews as a proxy to measure it [1, 4]. It has been concluded that weather affects both the frequency and the content of feedback provided by food consumers. Typically, sunny and pleasant weather leads to more frequent and more positive feedback, as low humidity and high sunlight are associated with high mood. At the same time, reviews written on rainy or snowy days, i.e. days with precipitation, tend to have lower ratings. While seasonal food consumption

[2] News portal of Latvian Radio and Latvian Television https://eng.lsm.lv/article/culture/food-drink/traditional-and-national-latvian-foods-whats-the-difference.a466155/

patterns are culturally specific and vary across geographic regions, weather-related preferences appear to be universal.

A large-scale study of demographics, weather and online reviews of restaurant recommendations shows that pleasant weather not only affects the content of the review, but also the frequency, which is higher than in less pleasant weather conditions [1]. This is an important indicator that a review can serve as a proxy for measuring the impact of the weather on mood, and thus on the food consumption experience. Consumer comments and word-of-mouth have also been studied in relation to the weather, suggesting that consumers' pre-consumption mood directly influences their post-consumption mood and, accordingly, their satisfaction with the service. Pre-consumption mood, in turn, is considered via weather conditions, with eight weather-related variables considered, including visibility, rain, storm, humidity, wind speed, barometric pressure and other variables. By including temperature, air pressure and rain as covariates, the researchers were able to reduce unexplained variance and improve the results of the experiment. This study successfully links weather to mood and its transfer to affective experience and consumer behaviour [4].

Reviews, word of mouth or tweets are language based proxies to determine attitude and related emotions towards the given topic, therefore, a deeper understanding of language and how we describe foods is a prerequisite to understanding the dynamics between the individual and the group when it comes to food choice. The language of how the dish is described matters, and the taste of the dish can change just because the wording of how food is described, has been changed [19]. Instead of language analysis of how particular foods are described, we focus on sentiment analysis which can be of great use for food language analysts to gain a more holistic view of how particular foods are discussed in different societies. With this research we aim to illustrate the utility of data coming from languages less resourced and less spoken. Thus, the LTEC, a unique resource devised for the analysis of Latvian food-related tweets, has been used in this research. It might serve as a pilot corpus for other less-resourced languages and contribute to a better understanding of the differences in food narratives depending on the language we use [20].

The interdisciplinary nature of food-related data poses challenges for the use of social network data. There are limitations to the fragmented nature of social media data: for example, Twitter users are digitally active and a relatively affluent part of the population, so results cannot be generalised to the whole of society. Nevertheless, even acknowledging the fragmentation, new research exploring the use of social media data can be of value to policy makers and those encouraging particular behaviours among food consumers. Another line of research that demonstrates the usefulness of social media analysis is looking at how digital food affects our analogue lives and eating behaviour in particular [21]. Correlating social media results about food in a particular region with sales data could be the next step in our analytical approach, as there is general agreement that digital content influences purchasing behaviour in our analogue lives, but this lacks granularity when it comes to exact correlations and proof of statements.

4. Data Collection and Processing

Our analysis examines the LTEC, which contains 2.4M tweets generated by 169k users. It has been collected for over 10 years by following 363 eating-related keywords in Latvian. The dataset provides some additional metadata about each tweet, such as location (when available), a list of food items mentioned in the tweet text, and a separate subset of tweets with manually annotated sentiment classes - positive, neutral and negative.

Since the corpus contains normalised versions of all food items in singular nominative form for each tweet, we used these to further select only the specific tweets for our analysis. This was done by firstly compiling a list of most used meat-related nouns (see Table 1), and then selecting only the very narrow subset which mentions either beef, chicken or pork.

liver	sausage	chop	bacon	roast	**chicken**	deer	bratwurst
beef	schnitzel	fillet	goose	gyros	ribs	ham	salami
steak	**pork**	cutlet	steak	lamb	meat	meatball	turkey

Table 1
The list of meat products included in our experiment.

4.1. Tweet Sentiment Analysis

We used the 5420 annotated tweets to fine-tune a pre-trained multilingual BERT [22] model for the sentiment analysis task along with ∼20,000 sentiment-annotated Latvian tweets from other sources[3] so that the model would generalise better. We evaluated the sentiment model on the 743 tweet test set provided in LTEC and reached an accuracy of 74.06%. Our result outperforms the best accuracy reported by the authors of LTEC, who used a Naive Bayes model on stemmed data and reached 61.23%. However, this was expected since they used ∼20% less training data, and BERT or other transformer-based models have outperformed previous state-of-the-art methods in many language processing tasks, including classification. We then used the model to automatically classify all tweets in LTEC as positive, neutral or negative for further analysis.

To verify the quality of the sentiment analysis model, we selected 50 random automatically classified tweets from each year between 2011 and 2020 and performed a manual evaluation. Twelve human evaluators were asked to individually judge each of the 500 predictions by the model and provide a suggested alternative sentiment class for cases where they deemed the model to be incorrect. We used the majority vote of the human evaluators as the correct answer in cases where they disagreed on a particular evaluation and considered two classifications as correct in the 21 cases where the majority opinion was split in half (for example, 6 positive and 6 neutral). The overall agreement of the evaluators was 70.48% with a free marginal kappa [23] of 0.56 (values from 0.40 to 0.75 are considered intermediate to good agreement). The accuracy of the model according to the majority of human evaluators on this set was even higher, reaching 86.40%, while the accuracy of the average human evaluator compared to the majority was only 80.25%. This shows that 1) the tweet texts are not always trivial enough to be unequivocally classified into just one of the three sentiment classes, and 2) the model is good enough to be used on the scale of the whole dataset.

[3] https://github.com/Usprogis/Latvian-Twitter-Eater-Corpus/tree/master/sub-corpora/sentiment-analysis#other-latvian-twitter-sentiment-corpora

4.2. Tweet Alignment with Weather Data

To conduct our analysis in relation to weather data, we used a combination of two data sources - the LTEC for tweets and weather data exported from Meteostat[4]. We mainly focused on tweets and weather relating to Riga, the capital of Latvia, since most tweets with location data originated there, and it was difficult to obtain detailed historical weather data for the smaller regions.

Among the tweets, 167k have location metadata specified, of which 68k were from Riga and 9k more from areas around Riga. To further increase the number of location-related tweets, we selected all remaining tweets which mention Riga or any of its surrounding areas (like Marupe, Kekava, Salaspils, Adazi, etc.) in any valid inflected form. This added 54k tweets, giving a total of 131,595.

From the Meteostat website, we could reliably obtain only data for temperature and precipitation, while data for snowfall was only available up to the end of 2017, and data for wind speed and air pressure was only available from July 2018 and onward. Figure 1 shows a visual depiction of the data gathered. There was no available data to trace daily sunshine directly, but it can be inferred from looking at precipitation, snowfall and air pressure.

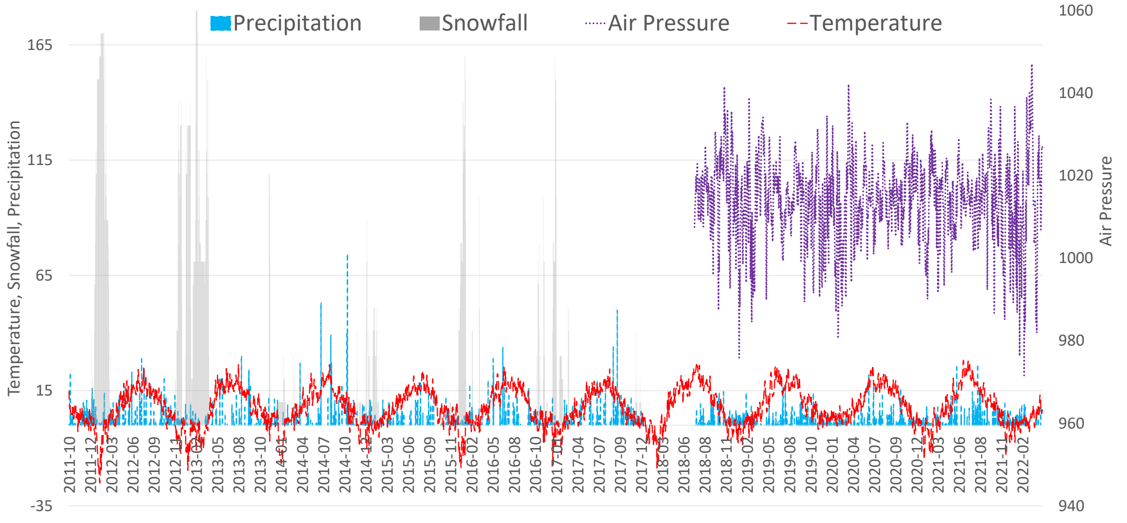


Figure 1: Visualisation of available weather data from Meteostat.

4.3. Limitations and Assumptions

Our work has several important limitations that can be grouped into the categories of 1) data availability, 2) demographic profile of the tweet author, and 3) generalisation of results. First, we were only able to obtain fairly superficial weather data, while subtleties such as weather changes during the same day were not taken into account due to the lack of such details. Second, we cannot provide a demographic perspective of the usual tweet author in LTEC, and our analysis includes

[4] https://meteostat.net/en/place/lv/riga

tweets from generally digitally literate people active on Twitter. Third, given the limitations discussed, our results are not an exact extrapolation of weather-related food perceptions in Latvian society. Nevertheless, our approach makes use of the growing LTEC and contributes to the understanding of the impact of weather on the part of Latvian society that tweets about food.

5. Analysis and Results

5.1. Food Tweet Relation to Type of Weather

While the results of tweet sentiment in terms of the percentage of negative, neutral and positive tweets are largely the same for all weather conditions, we can still observe significantly fewer positive tweets during windy and high pressure weather conditions, as can be seen in Table 2. We were surprised to see that even during low pressure weather conditions, tweets are not necessarily dominated by negative sentiment - on the contrary, food tweets were mostly associated with positive sentiment. This could be explained by the fact that people tweet about comfort food (e.g. coffee, chocolate, other) or that any food could be comforting during days of low pressure. This remains to be answered in a more fine-grained manual analysis.

	Cold	Warm	Windy	Snowy	Rainy	High Pres	Low Pres	Overall
-	12.59%	13.20%	23.15%	11.88%	13.63%	23.10%	12.63%	13.07%
0	37.25%	38.68%	**48.40%**	36.06%	38.64%	**48.26%**	38.72%	38.38%
+	**50.17%**	**48.12%**	28.45%	**52.06%**	**47.73%**	28.63%	**48.65%**	**48.55%**

Table 2
Weather relation to tweet sentiment. Rows -, 0, + specify negative, neutral and positive sentiment respectively.

The Table 3 shows that tea surpasses coffee in cold weather, and there is also a slight increase in tweets about chocolate in cold weather, while the frequency of ice-cream tweets doubles in warm weather. Interestingly, the number of tweets about meat, cake or soup in hot or cold weather remains broadly similar. While warm weather tweets include strawberries, cold weather tweets include gingerbread, which coincides with seasonal Christmas food. There are no other notable differences between warm and cold weather tweets, suggesting that spending most of our lives indoors has harmonised the foods we tweet about in different seasons and weather conditions.

Product	Tea	Coffee	Meat	Chocolate	Cake	Ice cream	Salad	Dumplings	Pancake	Sauce	Gingerbread
Rainy	8.78%	6.59%	4.20%	4.83%	**2.77%**	3.05%	2.19%	**2.25%**	**2.16%**	**2.01%**	1.49%
Windy	**6.64%**	5.94%	**9.44%**	**3.50%**	**4.20%**	**1.75%**	**3.15%**	**1.05%**	**0.70%**	**0.70%**	**2.10%**
Warm	7.70%	6.77%	4.38%	4.56%	2.85%	**4.04%**	2.14%	**2.28%**	2.07%	**2.07%**	**0.74%**
Cold	**10.08%**	6.73%	**3.95%**	**5.14%**	2.93%	2.39%	**1.81%**	2.12%	**2.20%**	1.65%	**2.10%**

Table 3
Comparison of top products during windy (wind speed $\geq$ 20km/h), rainy (precipitation > 0), cold ($\leq$ 0 °C), and warm weather ($\geq$ 0 °C).

A slightly different result can be seen in the Table 3 in relation to meat. It shows that in windy weather meat becomes the most popular food, while in rainy weather the results are similar to cold weather - where tea dominates. Although it is difficult to explain this result, it could be that wind is less visible than the temperature often reported in the media or precipitation, which can be seen before leaving the house, and therefore people may feel uncomfortably cold in windy weather without appropriate clothing, which could lead to a greater willingness to eat meat. Chocolate is twice as popular in rainy weather as in windy weather, and this could be related to the lack of sunshine in rainy weather, which needs to be compensated by chocolate, whereas a windy day can still be sunny.

5.2. Meat-related Tweet Analysis

Before turning to the results of the sentiment analysis of meat-related tweets, we first look at the distribution of sentiment across all food tweets in the dataset. In the tweets from 2011-2020, we can observe an overall decrease in positive sentiment and an increase in negative sentiment, as well as a comparatively large proportion of neutral tweets. Figure 2 shows the overall sentiment distribution over this period. It also shows that the number of positive tweets decreased until 2015, that the number of neutral tweets increased from 2015, and that the number of negative tweets increased from around 2018.

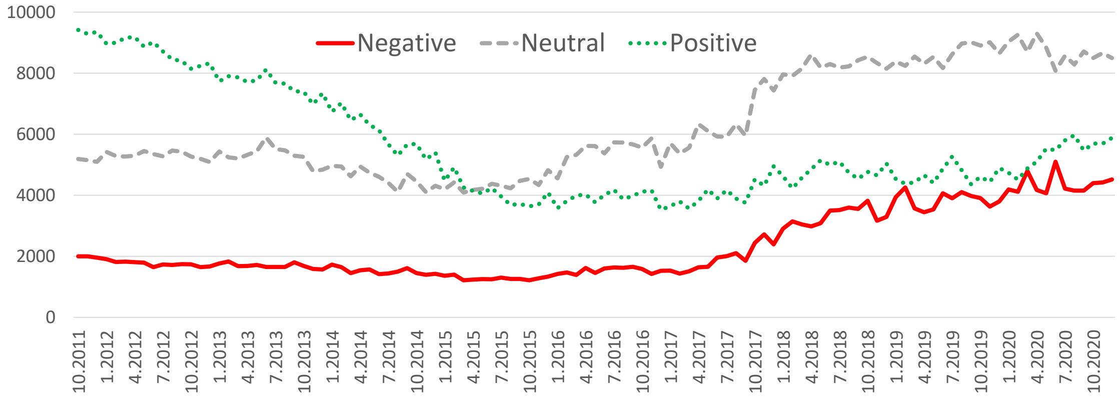


Figure 2: Distribution of overall tweet sentiment in LTEC over time from 2011 to 2020.

Figure 3 shows the sentiment over time of all meat-related tweets. In addition to selecting all inflections of the word "meat", we also include the specific meat products listed in Table 1 in this overview. Here we can see that until 2016, Twitter users were not overly active in discussions about meat overall. The proportion of tweets with neutral sentiments increased significantly between 2016 and 2018 and has remained largely stable since then, while the proportion of more polarised opinions - positive and negative meat-related tweets - still seems to be increasing slightly. Although the level of negative tweets is largely flat between 2011 and 2017, the one spike in March 2013 can be attributed to a scandal over the alleged use of horse meat in a popular butchery chain from Latvia[5]. Since 2016, tweets with negative sentiment have outnumbered those with positive sentiment, although the majority of meat-related food tweets can still be classified as neutral.

[5] https://www.theguardian.com/uk-news/2013/jul/19/horsemeat-scandal-meat-pies-latvia

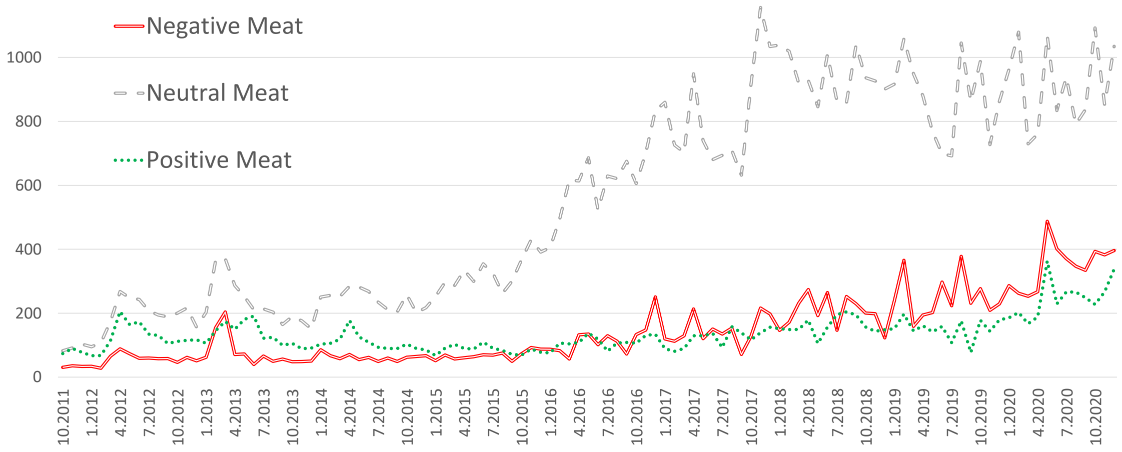


Figure 3: Temporal sentiment dynamics of meat-related tweets LTEC in 2011-2020.

To take a closer look at specific meat products, Figure 4 shows the differences in sentiment towards chicken, beef and pork. Again, we can see a sharp increase in neutral tweets from 2016, which could be explained by the rise in popularity of public lunch offers at local restaurants and other types of food-specific advertising. A neutral tweet in this case is a tweet that simply informs about the daily specials at a café or restaurant, without any emotional connotation to the food listed in those specials. However, with the onset of the Covid-19 pandemic, neutrality has given way to either positive or negative valence. One possible reason for this could be the closure of restaurants and other public spaces for food consumption, and correspondingly fewer such neutral lunch-offer-type tweets from the corporate sector.

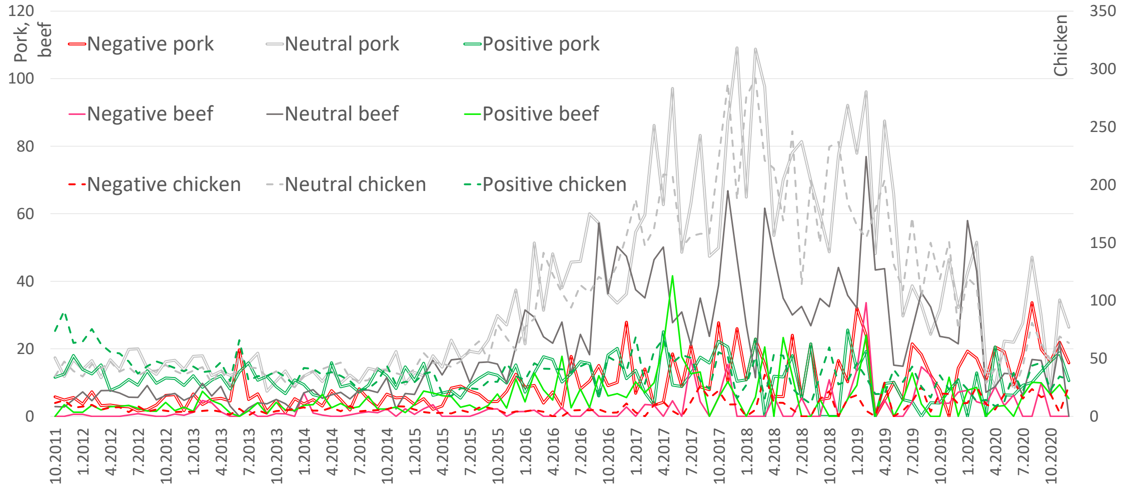


Figure 4: Temporal sentiment dynamics in 2011-2020: tweets mentioning beef, chicken or pork.

5.3. Vegans, Vegetarians and Alternative Proteins

In addition to posts directly mentioning meat-related products, we were also interested in whether meat alternatives are mentioned and how they are perceived on Twitter. Figures 5 and 6 give an overview of tweets mentioning either ’vegan’ or ’vegetarian’ in any inflection of the Latvian language or any inflection of the word ’protein’ in Latvian. Overall, we can observe a lower amount of tweets compared to those mentioning ’meat’, as well as a higher tendency of positive sentiment tweets. The dominance of positive sentiment tweets over neutral tweets, which dominate meat-related discourses, may mean that there are few tweets from, for example, the corporate sector in the form of lunch offers or sales of vegan/vegetarian food, or other marketing-related neutral tweets. Instead, as veganism and vegetarianism are not yet mainstream discourses, they are mostly discussed by people with strictly positive or negative attitudes towards them. Regarding negative sentiments, it should be noted that ’vegan’ is sometimes used as an insult in the Latvian Twitter space, referring to a person who is weak, incapable of doing activities that require physical strength, and does not have the work experience of younger generations. A new term, ’soy latte drinkers’, emerged in the debate when conscription was extended following Russia’s invasion of Ukraine. Young people protesting against conscription were ironically called ’soy latte drinkers’, implying their weakness due to their vegetarian or vegan lifestyle.

With regard to (alternative) proteins, we see a similar dynamic to that of vegan and vegetarian discourse, but with even lower frequencies, which means that the discussion about proteins in the Latvian Twitter space is very low, and when it takes place, it is mostly positive or neutral, with little informative content generated. These results of low frequencies mean that vegan and vegetarian diets and the search for alternative proteins remain marginal in the everyday discussions of Twitter users in Latvia.

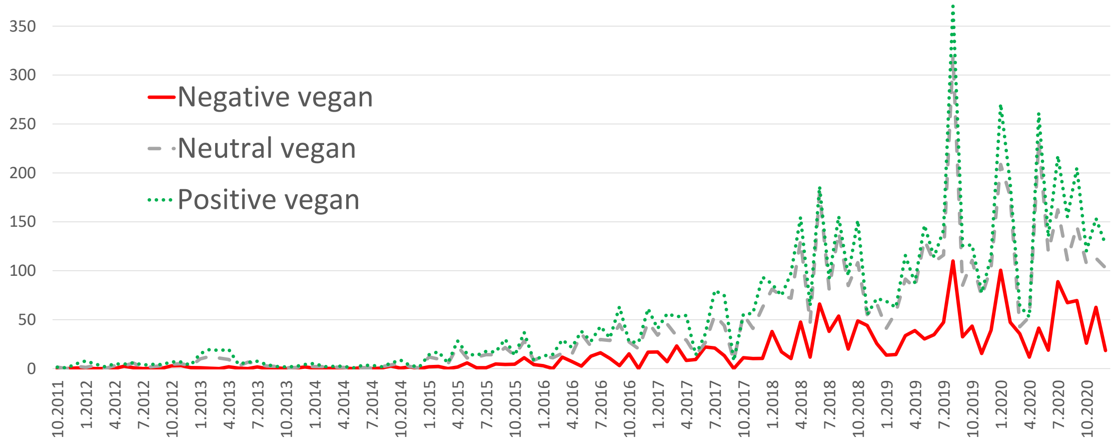


Figure 5: Distribution of tweet sentiment in LTEC over time from 2011 to 2020 of tweets mentioning vegan or vegetarian.

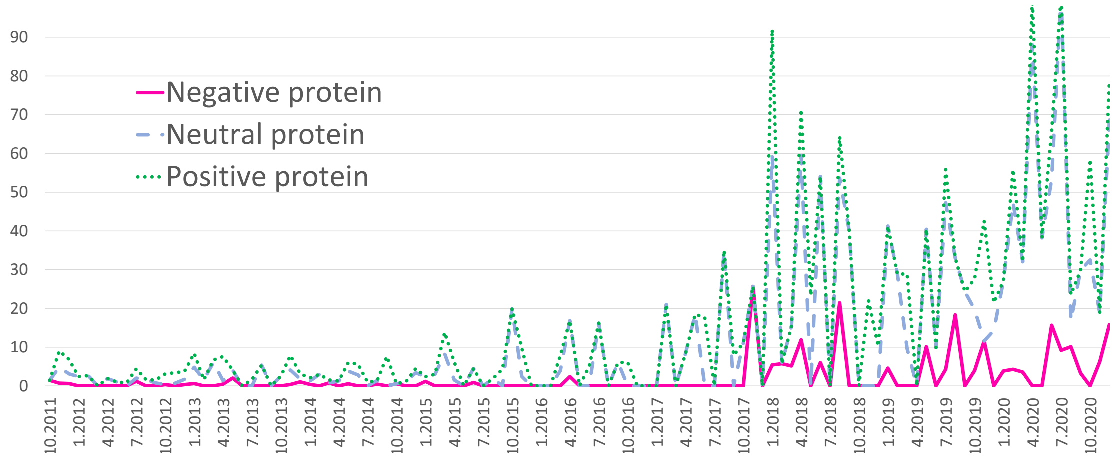


Figure 6: Distribution of tweet sentiment in LTEC over time from 2011 to 2020 of tweets mentioning protein.

6. Conclusion

Our analysis contributes to the understanding of how weather affects the mood of food consumers by showing that certain weather conditions, such as windy weather, affect the content of food tweets. This knowledge of tweet frequency and sentiment can be useful to public health policy makers and applied when nudging consumers to choose healthier food alternatives in different weather conditions and seasons. Recognising and understanding the impact of weather on food consumers and their affective responses helps to explain the complexities associated with food consumption - food waste, healthy vs unhealthy food choices and other issues.

We started with the statement that the future of food will be largely determined/dependent on the future of meat, as policy recommendations push for reduced meat consumption. This has paved the way for e.g. the development of alternative proteins, as more and more investment flows into this area, as well as the appreciation of vegetarian/vegan diets, which has come to shape the discourse in stark contrast to the discourse of meat lovers. In this case, social media can serve as a litmus test for public sentiment and attitudes towards meat consumption. Our research shows that negative sentiment towards meat is steadily increasing on Latvian Twitter, although neutral sentiment still dominates. The spread of the Covid-19 pandemic seems to have significantly reduced neutrality towards certain types of meat - chicken, beef and pork. All these data help us to track public attitudes towards meat consumption and to assess their willingness to change in the direction of the lower meat consumption future envisaged by policymakers. Looking at the sentiments and frequencies related to vegan/vegetarian food and protein, we conclude that there are not many discussions on Twitter related to these topics compared to meat.

These data can be useful for policymakers working with the public diets' shift towards more environmentally conscious choices. Knowing the dominating discourse in the society related to meat and being able to trace the sentiment changes over time, can potentially best signify the society's maturity for change as suggested by public health policymakers. These data can be useful also for industry players, such as retailers and meat producers who shape their own discourse on meat consumption in particular markets. For marketers, temporal sentiment dynamics related to meat are valuable sales and marketing data and can be utilised in their promotional activities.

Taking into account the seldom use of social media data in academic research due to the fragmented nature - user demographics unknown, data only from the relatively wealthy and digitally active part of society, particular preferences of the social network in focus - Twitter, while other different social networks also of use, we consider that our research provides important encouragement to utilise social network data. The utility of our research results can be seen via creating valuable insights into group dynamics of the particular society, and while fragmented and in many ways incomplete, social media data of a particular social network Twitter, can to a large extent impact individual food choices. Group dynamics of social media can signify and determine the trends that impact individual preferences and ultimately food choices. Therefore, when developing individual food and health applications, it is of utmost importance to include the context data of the individual, society, national cuisine, weather and seasonality in their analysis. Social media data serve to signify those various influential context factors as can be seen also from our analysis of a particular focus on meat consumption sentiments in the Latvian Twitter community.

To conclude, through our interdisciplinary research, we unravel the complex interplay between economics, digital platforms and practical knowledge within the Latvian food market and Twitter food posts. By understanding the impact of weather on consumer sentiment and tweet content, we provide valuable insights for policymakers to nudge consumers towards healthier choices. This highlights the potential of social media data to shape individual food preferences and drive societal trends, fostering a more sustainable and conscious food market in Latvia.

We plan to release the additional data and models generated in this research publicly. The automatically assigned sentiment classes will be added to the main corpus data repository on GitHub[6], and publish the sentiment analysis model to Hugging Face's model hub[7].

Acknowledgement

This research is supported by the project "Strengthening of the capacity of doctoral studies at the University of Latvia within the framework of the new doctoral model", identification No. 8.2.2.0/20/I/006

References

[1] S. Bakhshi, P. Kanuparthy, E. Gilbert, Demographics, weather and online reviews: A study of restaurant recommendations, in: Proceedings of the 23rd International Conference on World Wide Web, WWW '14, Association for Computing Machinery, New York, NY, USA, 2014, p. 443–454. URL: https://doi.org/10.1145/2566486.2568021. doi:10.1145/2566486.2568021.

[6] https://github.com/Usprogis/Latvian-Twitter-Eater-Corpus/

[7] https://huggingface.co/models

[2] C. Velasco, C. Michel, C. Spence, Gastrophysics: Current approaches and future directions, International Journal of Food Design 6 (2021) 137–152.

[3] C. Spence, Explaining diurnal patterns of food consumption, Food Quality and Preference 91 (2021) 104198. URL: https://www.sciencedirect.com/science/article/pii/S0950329321000252. doi:https://doi.org/10.1016/j.foodqual.2021.104198.

[4] M. Bujisic, V. Bogicevic, H. G. Parsa, V. Jovanovic, A. Sukhu, It's raining complaints! how weather factors drive consumer comments and word-of-mouth, Journal of Hospitality & Tourism Research 43 (2019) 656–681. URL: https://doi.org/10.1177/1096348019835600. doi:10.1177/1096348019835600. arXiv:https://doi.org/10.1177/1096348019835600.

[5] C. Spence, Explaining seasonal patterns of food consumption, International Journal of Gastronomy and Food Science 24 (2021) 100332. URL: https://www.sciencedirect.com/science/article/pii/S1878450X21000317. doi:https://doi.org/10.1016/j.ijgfs.2021.100332.

[6] M. Grivins, A. Halloran, M. Kāle, Eight megatrends in Nordic-Baltic food systems, Nordisk Ministerråd, -, 2020. URL: http://urn.kb.se/resolve?urn=urn:nbn:se:norden:org:diva-7127.

[7] M. Crippa, E. Solazzo, D. Guizzardi, F. Monforti, F. Tubiello, A. Leip, Food systems are responsible for a third of global anthropogenic ghg emissions, Nature Food 2 (2021) 1–12. doi:10.1038/s43016-021-00225-9.

[8] X. Xu, P. Sharma, S. Shu, T.-S. Lin, P. Ciais, F. N. Tubiello, P. Smith, N. Campbell, A. K. Jain, Global greenhouse gas emissions from animal-based foods are twice those of plant-based foods, Nature Food 2 (2021) 724–732. doi:10.1038/s43016-021-00358-x.

[9] T. Lancet, We need to talk about meat, The Lancet 392 (2018) 2237. URL: https://doi.org/10.1016/S0140-6736(18)32971-4. doi:10.1016/S0140-6736(18)32971-4, publisher: Elsevier.

[10] W. Min, S. Jiang, L. Liu, Y. Rui, R. Jain, A survey on food computing, ACM Computing Surveys 52 (2019) 1–36.

[11] M. Kāle, J. Šķilters, M. Rikters, Tracing multisensory food experiences on twitter, International Journal of Food Design 6 (2021) 181–212. doi:10.1386/ijfd_00030_1.

[12] C. Velasco, C. Michel, C. Spence, Gastrophysics: Current approaches and future directions, International Journal of Food Design 6 (2021) 137–152. doi:10.1386/ijfd_00028_2.

[13] U. Sproģis, M. Rikters, What Can We Learn From Almost a Decade of Food Tweets, in: Proceedings of the 9th Conference Human Language Technologies - The Baltic Perspective (Baltic HLT 2020), Kaunas, Lithuania, 2020, pp. 191 – 198.

[14] P. Puerta, L. Laguna, L. Vidal, G. Ares, S. Fiszman, A. Tárrega, Co-occurrence networks of twitter content after manual or automatic processing. a case-study on "gluten-free", Food Quality and Preference 86 (2020) 103993. URL: https://www.sciencedirect.com/science/article/pii/S0950329320302627. doi:https://doi.org/10.1016/j.foodqual.2020.103993.

[15] I. Croijmans, I. Hendrickx, E. Lefever, A. Majid, A. Van Den Bosch, Uncovering the language of wine experts, Natural Language Engineering 26 (2020) 511–530. doi:10.1017/S1351324919000500.

[16] LVĢMC, Climate of latvia, https://www.meteo.lv/en/lapas/environment/climate-change/

climate-of-latvia/climat-latvia?id=1471, 2009. Accessed: 2022-04-15.

[17] M. Kāle, M. Rikters, Fragmented and Valuable: Following Sentiment Changes in Food Tweets, in: Proceedings of Smell, Taste, and Temperature Interfaces Workshop, Yokohama, Japan, 2021, pp. 1 – 4.

[18] J. Maderer, A rainy day can ruin an online restaurant review, https://news.gatech.edu/news/2014/04/02/rainy-day-can-ruin-online-restaurant-review, 2014. Accessed: 2022-04-15.

[19] L. Bacon, J. S. Wise, S. Attwood, D. Vennard, The language of sustainable diets: A field study exploring the impact of renaming vegetarian dishes on u.k. café menus, 2019.

[20] A. Fenko, J. J. Otten, H. N. J. Schifferstein, Describing product experience in different languages: The role of sensory modalities, 2010.

[21] T. Andersen, D. Byrne, Q. Wang, How digital food affects our analog lives: The impact of food photography on healthy eating behavior, Frontiers in Psychology 12 (2021) 634261. doi:10.3389/fpsyg.2021.634261.

[22] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 4171–4186. URL: https://aclanthology.org/N19-1423. doi:10.18653/v1/N19-1423.

[23] J. J. Randolph, Free-marginal multirater kappa (multirater k [free]): An alternative to fleiss' fixed-marginal multirater kappa., 2005.

*August 22, 2022

Tracing the Proliferation of Socialist Realism Doctrine in Latvian Periodicals: Case Study of "Literature and Art" and "The Flag"

Anda Baklāne[1], Valdis Saulespurēns[2]

[1] *National Library of Latvia, 3 Mukusalas St, Riga, LV-1423, Latvia*
[2] *National Library of Latvia, 3 Mukusalas St, Riga, LV-1423, Latvia*

Abstract

The paper presents the results of a study on the dissemination of socio-political and aesthetic ideas of Socialist Realism doctrine in the Latvian periodicals *Literature and Art* (*LitArt*) and *The Flag*. Several programmatic, ideologically saturated articles that were published in *The Flag* in the 1940s were compared to the rest of the corpus to explore the proliferation and persistence of similar ideas in the course of following decades. Authors have employed methodologies commonly used for plagiarism detection: fingerprinting and the comparison of document similarity based on word embeddings and document similarity measures. In particular, three perspectives were used to examine the similarity and reuse of texts: comparison of matching 5-grams processed by the winnowing algorithm and comparison of documents based on the TF-IDF and Doc2Vec embeddings and cosine similarity metrics. To facilitate the analysis, the results were loaded in the open-source version of Neo4j graph database. The findings were further explored and evaluated qualitatively to identify the distribution of direct citations, frequently reused phrases and most similar documents.

Keywords

string similarity, document similarity, Latvian historical newspapers, Socialist Realism, discourse analysis

1. Introduction

This paper is part of a series of case studies aimed at researching the possibilities for implementing text similarity detection methodologies for the analysis of the collection of digitized historical newspapers of the National Library of Latvia.[1] Methods for identifying similarity in text documents have a broad range of applications, such as the detection of plagiarism [1] and studying the reuse of texts in historical newspapers [2] [3]. In this article, the authors investigate the usability of computational string and document similarity detection [4] [5] [6] in

DHNB2023 | Sustainability: Environment - Community - Data. The 7th Digital Humanities in the Nordic and Baltic Countries Conference. Oslo – Stavanger – Bergen, Norway. March 8–10, 2023.

anda.baklane@lnb.lv (A. Baklāne); valdis.saulespurens@lnb.lv (V. Saulespurēns)
https://lnb.lv/en/ (A. Baklāne); http://tuvenieks.lnb.lv:7474/browser/ (V. Saulespurēns)
0000-0002-0301-2504 (A. Baklāne); 0000-0002-9665-0125 (V. Saulespurēns)

DHNB Publications, DHNB2023 Conference Proceedings, https://journals.uio.no/dhnbpub/issue/view/875

[1] In the earlier case study, the methodology of LDA topic modelling was explored; see: A. Baklāne, V. Saulespurēns, The application of latent Dirichlet allocation for the analysis of Latvian historical newspapers: Oskars Kalpaks' case study. Nauka, tehnologii, innovacii, 2022, No.21, pp. 29-37; A. Baklāne, V. Saulespurēns, Latento Dirihlē sadalījumu modeļa izmantojums laikraksta *Latvijas Kareivis* tematu analīzē: Oskara Kalpaka gadījuma izpēte, Letonica, 2022, No.47, pp. 150-166.

recognizing the dispersion of similar discourse in the corpus of historical newspapers. The case study is constructed as an exploration of the proliferation of the lexicon and utterances related to the discourse of aesthetical tenets of Social Realism in two Latvian cultural periodicals – *The Flag* and *Literature and Art* (hereafter *LitArt*).

In the field of discourse analysis, the examination of discourse is understood as the study of texts and other semiotic phenomena taking into account a wide range of semiotic features and especially paying attention to the construction of meaning above the level of a sentence. In addition to that, discourse analysis is not a formal and purely linguistic analysis of a text and other semiotic entities: it always involves studying language in the context of society, culture, history, institutions, identity formation, politics, and power [7]. Depending on the context, discourse can be understood as (a) meaning-making as an element of the social process; (b) the language associated with a particular social field or practice (e.g. 'political discourse'); (c) a way of construing aspects of the world associated with a particular social perspective (e.g. a 'neo-liberal discourse of globalization') [7] [8]. When encoded in language and text, a particular discourse can manifest itself as a characteristic vocabulary and a set of stylistic and rhetorical devices that can be identified with different positions or perspectives of societal groups or actors. Discourse can be expressed in all dimensions of a text: semantic, syntactic, and pragmatic. In this case study, authors are focusing on the semantic aspects of the discourse by studying lexical features of texts on a phrase level, as well as looking at the overall similarity of documents based on word embeddings.

Socialist Realism is an artistic style and an aesthetic doctrine that was officially sanctioned and prevalent in the Soviet Union from 1932 to the mid-1980 [9]. According to the tenets of Socialist Realism, artworks should represent the objective reality and be realistic in style (all non-realist modernist styles are condemned as formalism), typical (present condensed ideal types of characters, classes, and circumstances), optimistic and progress-oriented, pedagogical, and profess loyalty to the Communist Party and its goals [10]. The introduction and dismissal of socialist realism did not occur simultaneously in all Socialist countries (e.g., in Latvia, it was fully implemented after the Second World War when Latvia was occupied by the USSR for the second time, not in the 1930s). Although in the USSR Socialist Realism officially remained a leading aesthetic standard and ideal until the 1980s, it was contested at different times in different countries throughout the 1950s and 1960s [11]. Hence, it is relevant to explore the temporal dynamics of the discourse.

The possibility of tracing the proliferation of Socialist Realism doctrine is a theoretical problem in itself that cannot be fully tackled in this article. On the one hand, the media discourse in the USSR is believed to be characterized by parrot-like repetitions of ideas and phrases pertaining to the doctrine. On the other hand, it has been often emphasized that the development of national literary traditions is a complex process and it is not always easy to discern the direct influence of Socialist doctrine from other influences and personal beliefs of writers and literary critics [12] [13]. What is the vocabulary of Socialist Realism? Which phrases are signaling the presence of the discourse? Complete formalization of the Socialist Realism discourse for to purposes of fully automated analysis, if possible, is beyond the scope of this case study. Instead, the authors of the paper have assumed a less ambitious approach: to look at the reuse of vocabularies and phrases through the lens of individual notable articles; quantitative analysis is supported by the qualitative evaluation of the results.

Table 1
Corpus statistics

Parameter	The Flag	LitArt	Total
Issues	607	2573	3180
Articles	39 406	79 047	118 453
Raw token count	62 835 625	67 098 255	129 933 464
Clean lemma count	48 324 759	52 051 659	100 376 418

Various approaches of computational text analysis can be employed to pursue the task of studying lexical features of a discourse, for instance: searching for specific pre-selected words and phrases in the corpus, identifying topics by means of topic modelling, measuring the similarity of documents (i.e., TF-IDF, Jaccard distance, Cosine distance, Hamming distance, Levenstein distance, or other), or searching for direct re-publications of texts, citations, and uncited reoccurrences of passages.

In the examination of *The Flag* and *LitArt*, authors have explored two approaches: (1) the analysis of matching 5-grams, based on the application of document fingerprinting performed by the winnowing algorithm [5] [14]; (2) the analysis of document similarity, based on TF-IDF [15] and Doc2Vec embeddings [16] that both were further compared by Cosine similarity [17]. Each of the measurements explored highlights a different aspect of the similarity of texts: the comparison of 5-grams is geared towards identifying the reuse of phrases, not taking into account the overall similarity of documents; measuring the cosine similarity between documents, on the other hand, allows to trace the overall similarity of documents that may stem from the similarity of topics or similarity of the vocabulary preferred in a particular discourse, or both.

2. Corpus and seed articles

Two titles of periodicals were selected for the case study – monthly literary magazine *The Flag* (*Karogs*, 1940-1995) and the weekly newspaper *Literature and Art* (*Literatūra un Māksla*, hereafter: *LitArt*, 1945-1994) that covered a broad range of topics related to literature, visual arts, and architecture. *LitArt* was a literary, artistic, and political weekly newspaper of the creative unions of the Latvian Socialist Republic (LSSR); *The Flag* was a monthly literary magazine published by the Writers' Union of the LSSR. Both periodicals were initiated in the 1940s and retained their status as the most important sources of information on current events in literature, art, and architecture until the 1990s (for corpus statistics see Table 1).

The corpus was derived from the digital collection of periodicals of the National Library of Latvia[2]. During the digitization, the content of periodicals has been segmented on the level of individual articles; in most cases, the titles and authors of the articles were also detected and marked providing additional features for further in-depth analysis. The corpus was lemmatized

[2] Access to the digitized periodicals is provided through the sites http://periodika.lv/ and https://lndb.lv/; corpora used in this study are protected by copyright and can be accessed for research purposes on demand: https://dom.lndb.lv/data/obj/1282468.html; https://dom.lndb.lv/data/obj/1282469.html

with the Latvian natural language processing tool pipeline NLP-PIPE[3] [18] and stopwords were removed. To mitigate the effects of optical recognition errors, all one-symbol words were removed, as well as words not recognized by the lemmatizer.

To construct the case study, five articles (hereafter: 'Seeds') were selected to represent the discourse of Socialist Realism:

- Seed 1: Upīts, A., "Marksistiska literatūrzināte un kritika. 1. d. Pretī jauniem apvāršņiem (Marxist literary studies and criticism. Part 1: Towards the new horizons)", Karogs, No.1 (1940), pp. 97–100.
- Seed 2: "Padomju Latvijas rakstnieku deklarācija (Declaration of the writers of Soviet Latvia)", Karogs, No.3 (1940), pp. 323–324.
- Seed 3: Vipers, B, "Sociālistiskais reālisms mākslā (Socialist realism in art) ", Karogs, No.3 (1940), pp. 437–442.
- Seed 4: Pupa, A., "Latvijas PSR tēlotājas mākslas izstāde (Visual arts exhibition of Latvian USSR)", Karogs, No.1 (1941), pp. 86–92.
- Seed 5: Rokpelnis, F., "Cīruļi sauc cīņā (Larks are calling to battle)", Karogs, No.1 (1943), pp. 84-86.

The seed articles were selected manually aiming to include articles that, based on a subjective judgement of the authors of this paper, contain ideologically charged Socialist vocabulary. In addition to that, the selection was designed to include articles according to the following criteria: (1) articles related to both literature and visual art; (2) theoretical writings as well as practical criticism; (3) works of renowned authors who were more likely to be cited in the following years. Since the scope of the analysis did not entail systematical research of the republishing patterns between *The Flag* and *LitArt* and the sample was very small, all seed articles were selected only from the early editions of *The Flag*.

Seeds 1, 2, and 5 are pertaining to the field of literary studies, 3–4 are related to visual art. Seed 2 – "Declaration of the writers of Soviet Latvia" – is a short manifest that itemizes the aesthetic principles of Soviet writers and pledges loyalty to V.I. Lenin, J.V. Stalin, the Communist Party, and the Soviet People. Seeds 1 and 3 are programmatic theoretical accounts of the state and principles of Socialist art written by prominent scholars in their respective fields – literary theory and criticism (Andrejs Upīts) and art history and theory (Boriss Vipers). Seeds 4 and 5 are examples of art criticism: Seed 4 provides an account of the first art exhibition organized in the Latvian USSR during the first Soviet occupation (1940-1941), while Seed 5 is a literary review that discusses a poetry book of a notorious Soviet poet Jānis Sudrabkalns titled "Larks Are Calling to Battle".

3. Methodology

Matching of identical and near-identical text strings and document similarity measures were used as a proxy to recognize the distribution of discourse similar to the discourse represented in the seed articles. After the cleaning and lemmatization, the corpus was processed to acquire

[3] NLP-PIPE: Latvian NLP Pipeline as a Service. Accessible: http://nlp.ailab.lv/

Table 2
Top 10 of sorted n-grams in the Corpus

3-grams	4-grams	5-grams
to be, which, that	highly, worker, art, accomplishment	supreme, latvia, council, presidium, ssr
which, about, that	to be, which, about, that	highly, worker, art, accomplishment, ssr
time, same, that	latvia, soviet, writer, union	highly, worker, lssr, art, accomplishment
also, to be, that	supreme, latvia, council, ssr	supreme, honorary, council, presidium, diploma
to be, about, that	great, october, revolution, socialist	anniversary, great, october, revolution, socialist
to be, that, all	highly, latvia, accomplishment, ssr	supreme, honorary, council, presidium, ssr
to be, something	supreme, council, presidium, ssr	highly, latvia, art, accomplishment, ssr
to be, that, or	but, time, same, that	literature, museum, art, rainis, history
to be, that, he	about, same, understand, self	academic, ballet, opera, theater, state
but, to be, that	highly, artist, accomplishment, stage	award, supreme, latvia, council, ssr

three types of embeddings (vector representations of text strings) that would enable the computational comparison of documents: (1) 5-grams processed by the winnowing algorithm, (2) TF-IDF embedding, and (3) Doc2Vec embedding. To further map the similarity of documents two distance measures were used: (1) Jaccard similarity for 5-grams and (2) cosine similarity for tf-idf and Doc2Vec embeddings. To make the similarity more accessible to the qualitative analysis, matrices of the embeddings were loaded in the open-source version of Neo4j graph database.[4] Finally, the results of the quantitative analysis were explored, evaluated, and interpreted qualitatively, zooming in to individual cases of matching 5-grams and clusters of most similar documents. In this process, the interactive Noe4j and Plotly visualizations as well as various lists of most common n-grams were immensely helpful to perform a closer inspection: a small selection of these figures and tables is included in this article.

Following Python libraries and other tools were used to perform the computations:

- Pandas[5] – data wrangling;
- Numpy[6] – numerical calculations;
- Gensim[7]- Doc2Vec;
- Scikit-learn[8] – tf-idf, similarities;
- Plotly[9] - visualizations;
- Neo4J drivers[10] for Python.

The **winnowing algorithm** is a variation of a fingerprinting algorithm that converts text documents into a set of hash values called "fingerprints"; the winnowing algorithm is proven to be efficient for finding matches of a set length of text strings (n-grams) [14].

[4] Neo4j graph database: https://neo4j.com/; local application in the server of the National Library of Latvia: http://tuvenieks.lnb.lv:7474/browser/

[5] Pandas - open source data analysis and manipulation tool: https://pandas.pydata.org/

[6] NumPy - fundamental package for scientific computing in Python: https://numpy.org/doc/stable

[7] Gensim - Python library for representing documents as semantic vectors: https://radimrehurek.com/gensim/

[8] Scikit-learn - Python module integrating a wide range of state-of-the-art machine learning algorithms for medium-scale supervised and unsupervised problems: https://jmlr.csail.mit.edu/papers/v12/pedregosa11a.html

[9] Plotly Graphing Library - https://plotly.com/python/

[10] Neo4J - NoSQL graph database: https://neo4j.com/docs/api/python-driver/current/

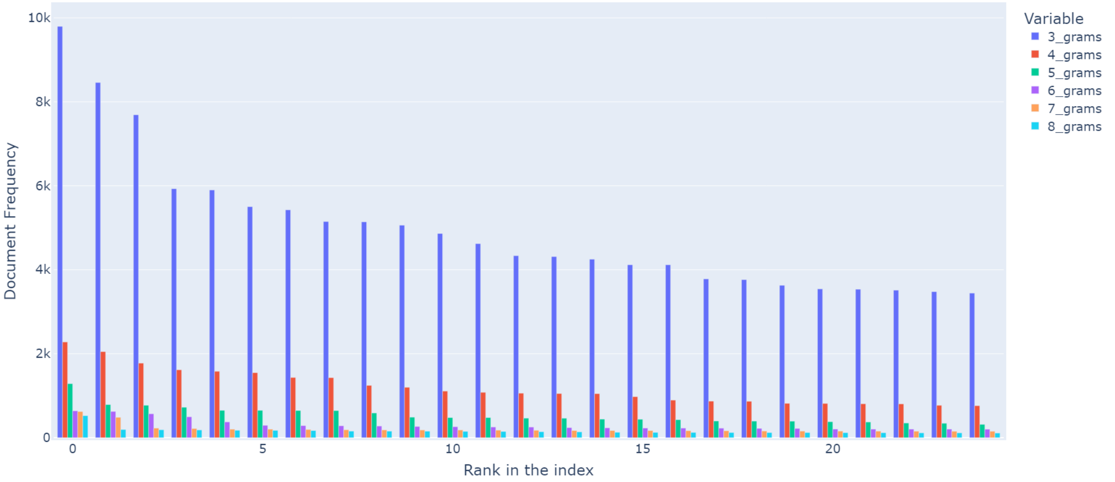


Figure 1: Comparative document frequency of n-grams (n is 3 to 8)

For the **5-gram embeddings**, a dictionary of n-grams of size 5 was calculated for each document of the corpus before applying the winnowing algorithm; the words in the 5-grams were alphabetically sorted, hence allowing for more flexibility in finding ma tches. After the pre-processing of the corpus, all text strings consisting of 5 consecutive words matching the strings found in seed documents were identified in the corpus; the instances of the matching strings were analyzed qualitatively and visualized by using the Python Plotly library.

The length of n-grams was selected based on the explorative analysis of the document frequency of n-grams (where n = 3-8) in all documents of the corpus (Figure 1). Subjective inspection of 1000 most common reoccurring 3-grams revealed that this index does not include phrases immediately useful for qualitative analysis of the discourse of socialist realism - this data entails mainly common combinations of words (e.g., "which also is", "all this is", "he is the one", "it is about" etc.) and some mentions of organisations and events, such as "Great Patriotic War" (13th most common 3-gram), "Opera [and] Ballet Theatre" (27th most common 3-gram), "First World War" (67th most common 3-gram) etc. The indices of 1000 most common n-grams with the count of n>3 contain a much larger proportion of the mentions of named entities - organisations, events, and honorary titles of awarded writers and artists (e.g., "People's Artist [of the] Latvian SSR", "Highly Accomplished Artist [of the] Latvian SSR"). The analysis suggested that n-grams with n>3 contain information potentially relevant to the studying of the discourse, with 4-grams most prominently featuring complete or partial mentions of named entities and 5/8-grams increasingly revealing additional elements of the discourse. For 'n' larger than 4, the reoccurrence of phrases in the Corpus drops significantly (Figure 1), however, the amount of reoccurring 4-grams is large also due to the overlap of two or three 4-grams that are layered upon the named entities that are more than 4 words long (see table 2 for top 10 most common 3/5-grams). Thus, 5-grams were selected for the case study as a middle ground;

after initial inspection, 5-grams were assumed to represent a unit of language that is suitable not only for quantitative analysis but for subjective qualitative evaluation as well which was a relevant aspect within the framework of this study.

TF-IDF and Doc2Vec embeddings were used to prepare documents for further similarity measurments. **TF-IDF or "term frequency - inverse document frequency"** is a measure of word frequency in the documents where the importance of a term is inversely related to its frequency across documents in the whole corpus. Namely, after counting the frequencies of all words, the overall importance of words that appear in the majority of documents in the corpus is reduced - based on the premise that the most frequently used words are functional words that do not carry essential information about the subject discussed in a given document [15].

$$\text{TF-IDF}(t, d, D) = \text{TF}(t, d) \cdot \text{IDF}(t, D) \tag{1}$$

- TF(t, d): Term Frequency, the number of occurrences of term t in document d
- IDF(t, D): Inverse Document Frequency, calculated as

$$\text{IDF}(t, D) = \log \frac{N}{\text{DF}(t, D)} \tag{2}$$

- t: the term or word of interest
- d: the document in which the term t appears
- D: the collection of documents
- N: the total number of documents in the collection D
- DF(t, D): Document Frequency, the number of documents in the collection D containing the term t

For the TF-IDF embedding, a word vector consisting of the top 2000 word lemmas was generated for each document. The documents were further compared by using the metrics of cosine similarity. **Cosine similarity** is a type of hierarchical agglomerative clustering, the normalized dot product between two vectors representing the compared document [15, 17] that is proven to be efficient for the comparison of text documents [6]. The results were analyzed qualitatively and visualized by using Neo4j network dependency graphs.

Doc2Vec is an unsupervised learning algorithm that represents each document by a dense vector which is trained to predict words in the document [16]. In contrast to the TF-IDF-based embedding, Doc2Vec creates vector representations not for de-contextualised individual words but for sentences. For the Doc2Vec embedding, a size 50 vector was generated for each document after training the corpus for 20 epochs (i.e., repetitions of the training process; each epoch represents one full iteration over all of the training data). The hyperparameters used were the default parameters suggested by the gensim library and based on the experiences of the original implementers of Doc2Vec. The results were further analyzed qualitatively and visualized by using Neo4j network dependency graphs.

Each of the approaches - the analysis of 5-grams and the comparison of the similarity of documents - provide a different perspective on the similarity of the articles of the corpus, allowing one to reflect on different aspects of the proliferation of discourses.

Table 3
Number of documents per number of matching 5-grams in each seed document

Seed 1	Seed 2	Seed 3	Seed 4	Seed 5
172: 1 doc.	73: 1 doc.	40: 1 doc.	67: 1 doc.	43: 1 doc.
30: 1 doc.	68: 1 doc.	8: 1 doc.	28: 1 doc.	17: 1 doc.
26: 1 doc.	46: 1 doc.	2: 23 docs.	3: 2 docs.	13: 1 doc.
5: 1 doc.	31: 1 doc.	1: 544 docs.	2: 21 docs.	11: 1 doc.
2: 2 docs.	30: 1 doc.		1: 212 docs.	8: 1 doc.
1: 69 docs.	29: 1 doc.			5: 5 docs.
	14: 1 doc.			2: 5 docs.
	13: 1 doc.			1: 88 docs.
	4: 15 docs.			
	3: 23 docs.			

4. Results

4.1. Matching 5-grams

Comparison of 5-grams is focused on identifying identical or near-identical text strings. In this case study, the method proved to be especially useful for detecting direct citations and commonly used phrases (at the same time, it does not account for the broader lexicon and topics of the articles).

The analysis of the 5-grams contained in the five seed articles demonstrated that all instances of 5-gam matches consisting of five or more than five 5-grams were direct quotes of the seed articles. Matches that entailed 1 to 4 5-grams in the majority of cases contained phrases that were present in the seed articles, however, were not used as direct citations from those articles. These phrases broadly fell into four categories:

- ideological aphorisms directly related to SR doctrine - expressions attributed to V.I. Lenin and J.V. Stalin ("nationalist in form, socialist in content");
- other idioms, stock phrases with no authorship ("nation's best sons and daughters");
- named entities ("communist party", "soviet union");
- other common multi-word expressions ("it is self-evident that").

According to the analysis of the matching 5-grams (Table 3), longer direct citations consisting of 5 or more consecutive 5-grams were relatively rare: 2 to 8 cases of text re-use in later publications. For instance, the "Declaration of the writers of Soviet Latvia" (Seed 2) was cited in eight other articles in*The Flag* and *LitArt* where the length of the cited passage varied from 13 to 73 consecutive 5-grams (see Table 3). All citations occurred in the event of the anniversary of the founding of the Writers' Union of the Latvian USSR. Similarly, the review of the first Soviet art exhibition in the Latvian USSR (Seed 4) was directly cited two times - on the dates when the exhibition was commemorated in later years. In the case of the literary review of the poetry collection "Larks are calling to battle" (Seed 5), the cases of the text re-use were not

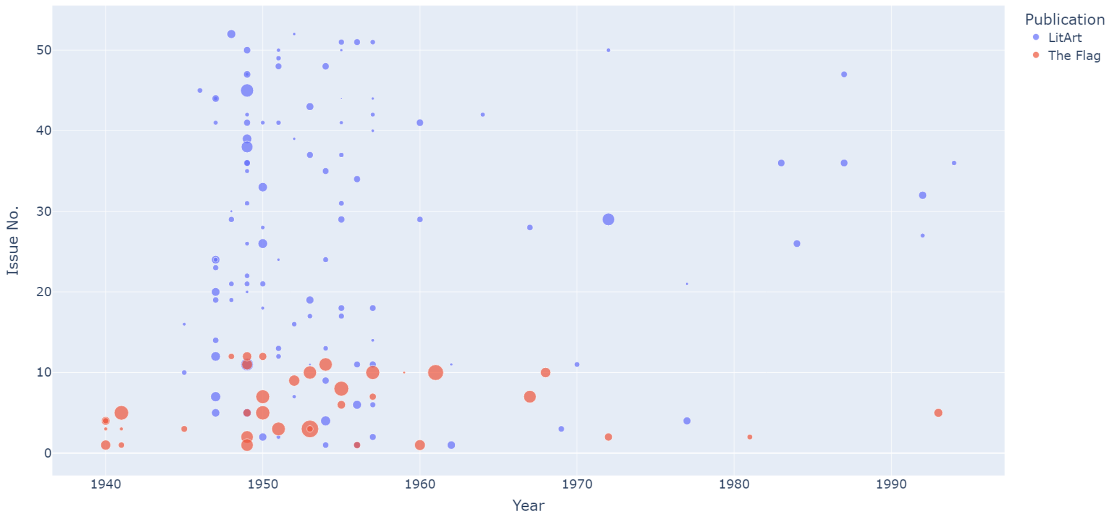


Figure 2: Most popular matching n-grams: "nationalist in form socialist in content"

direct citations of the words of the critic but the citations and re-publications of the lines of Jānis Sudrabkalns's poems that were also cited in the seed article.

In the analysis of the reuse of individual phrases, two doctrine-related 5-grams stood out among other idioms.

Firstly, it was the phrase containing the words "nationalist in form, socialist in content", found in Seed 2 - "Declaration of the Latvian Soviet Artists". In the Declaration, we find the following lines: "The work of the Soviet man, his high socialist moral, unshakeable Bolshevik stance, and submission to the communist idea – i.e. Soviet life in its entirety – needs to be vividly represented by writers in their works. These works need to be *Socialist in content and nationalist in form*." In other texts of the Corpus, the phrase "nationalist in form, socialist in content" was sometimes attributed to J.V. Stalin but sometimes identified as a commonly known and accepted truth. Outside the *Flag* and *LitArt* Corpus, in the body of Stalin's work, the phrase appears in Stalin's speech "On the Draft of Constitution of the U.S.S.R." where we find it in the following context: "The absence of exploiting classes.., the fact that power is in the hands of the working class.., and, finally, the flourishing national culture of the peoples of the U.S.S.R., *culture which is national in form and Socialist in content* - all these and similar factors have brought about a radical change in the aspect of the peoples of the U.S.S.R.[11]". Figure 2 demonstrates the distribution of the reoccurrences of the phrase in the course of time (each bubble represents one occurrence in an article; the size of a bubble represents the length of the article).

Another phrase that stood out as a notable signal of the ideological discourse was the adage "feelings, thoughts and will of the masses" that was used in the art review "Visual arts exhibition of Latvian USSR" (Seed 4). The author of the review proclaims: "Art belongs to the people. It

[11] J.V. Stalin, On the Draft of Constitution of the U.S.S.R., Works, Vol. 14, Red Star Press Ltd., London, 1978

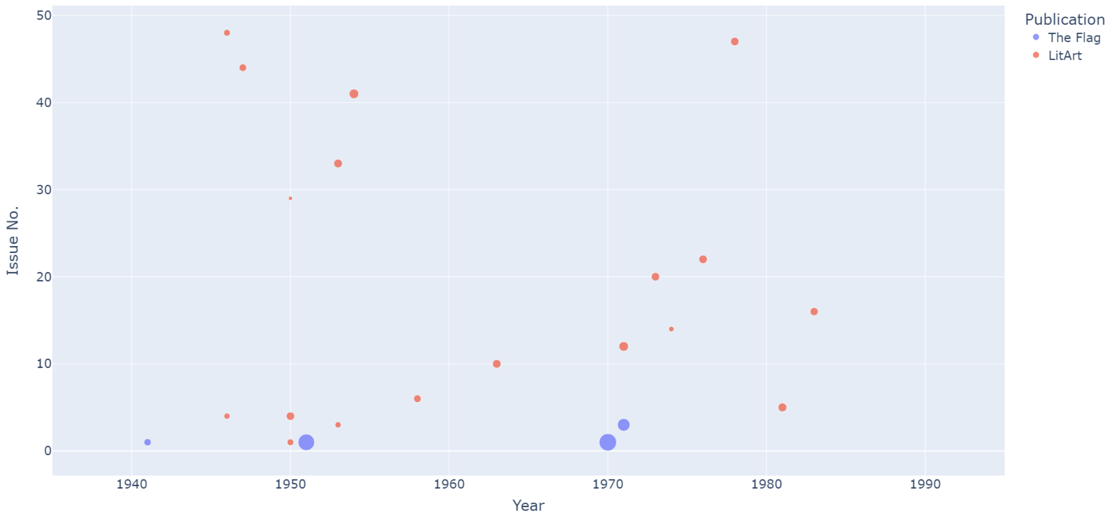


Figure 3: Most popular matching n-grams: "feelings, thoughts and will of the masses"

must with its widest stretching roots go out into the very thick of the broadest masses. It must combine the *feelings, thoughts and will of the masses and uplift them*." In this article and all later cases of the text re-use in the Corpus, the phrase is attributed to V.I. Lenin; although it is one of the most popular Lenin's aesthetic ideas, it is not sourced from the writings of Lenin himself but is known from the recollections of Clara Zetkin[12] (see Figure 3 for the distribution of occurrences over time).

In addition to phrases that could be directly traced back to the discourse promoted by Soviet ideologues, there were idioms that signaled the presence of an ideologically inclined discourse more subtly. For instance, the phrase "nation's best sons and daughters" could not be attributed to a particular author or source, however, it was recognizable as a rhetorical device, used to persuade and influence readers' opinions on the matter at hand. In the examples recovered in this case study, "nation's best sons and daughters" were usually people who have suffered in the name of the Socialist revolution and Communist future. Initially used in the "Declaration of Writers of Soviet Latvia", the phrase mostly appears in the 1950s (see Figure 4). It is interesting to note that in the use cases in the 1990s (after the regaining of Latvia's independence), in one instance, the phrase is used ironically, ridiculing the Soviet discourse, while in the second example, we find the same rhetorical device that is now employed to support another (Latvian nationalist) discourse - to reference people who have suffered from the Soviet deportations.

The largest number of re-used 5-grams, nevertheless, fall into the category of common multiword expressions, such as "it is self-evident that", phrases that contain words "it can be said that..." and "consequently...". In contrast to the ideologically significant phrases that are mostly used in the 1950s, common phrases are evenly distributed throughout all years up until the

[12] C. Zetkin, Reminiscences of Lenin (January 1924), Reminiscences of Lenin, International Publishers, 1934.

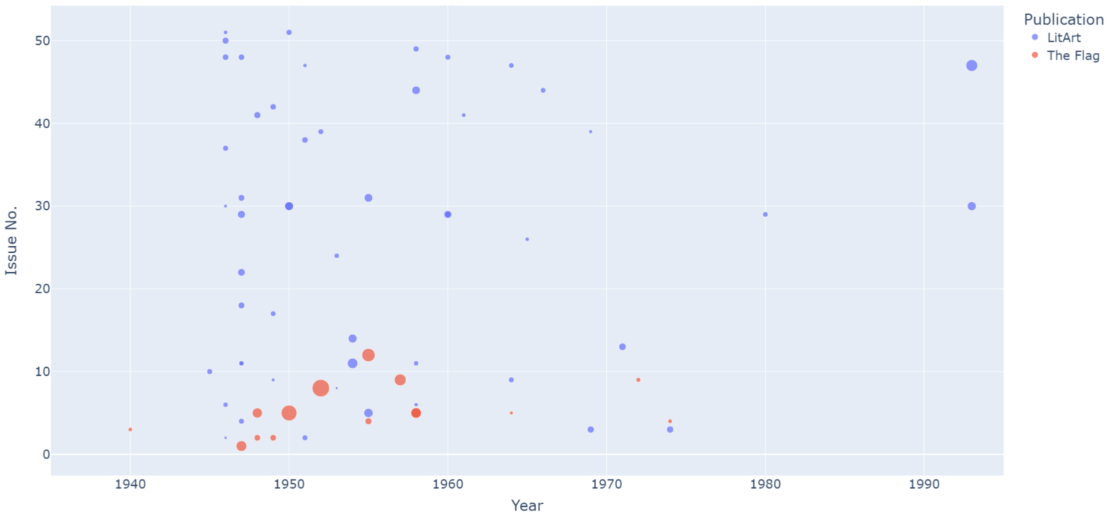


Figure 4: Most popular matching n-grams: "nation's best sons and daughters"

1990s (see Figure 5).

Hence, although the analysis of 5 seed articles is not nearly sufficient to generalize about overall trends, so far, looking at the matches of 5-grams of a sample, we can support the hypothesis that frequent reiteration of the doctrine-related rhetoric was typical for the writings of the 1950s and by the 1960s was already declining. The analysis of a larger number of seeds, including texts outside the current corpus, could bring more evidence to test the claim.

It is relevant to notice that rhetorically significant phrases could be shorter than those consisting of five consecutive words. Already among the titles of the seed articles, we find a high-flown three-word phrase "towards new horizons" that is frequently found in the corpus and could be studied as a signifying element of a discourse. However, as shown in the preliminary examination of most common n-grams, the index of 3-grams contains a very large proportion of commonly used word combinations, hence, it is not particularly suitable for approaches where subjective qualitative inspection is involved. However, in other types of approaches, 3-grams should be also considered.

4.2. Document similarity

In contrast to the 5-grams, the TF-IDF and Doc2Vec embeddings are more focused on the overall similarity of the lexical content of the articles and are not as useful for identifying the reuse of specific text strings (Table 4).

The TF-IDF and Doc2Vec embeddings each provide a slightly different perspective on lexicons of articles: TF-IDF is based on the frequency counts of de-contextualized words while Doc2Vec embeddings are learned by the algorithm taking into account the context of sentences. The results of all three similarity metrics (5-grams, TF-IDF, Doc2Vec) are partially overlapping. Table

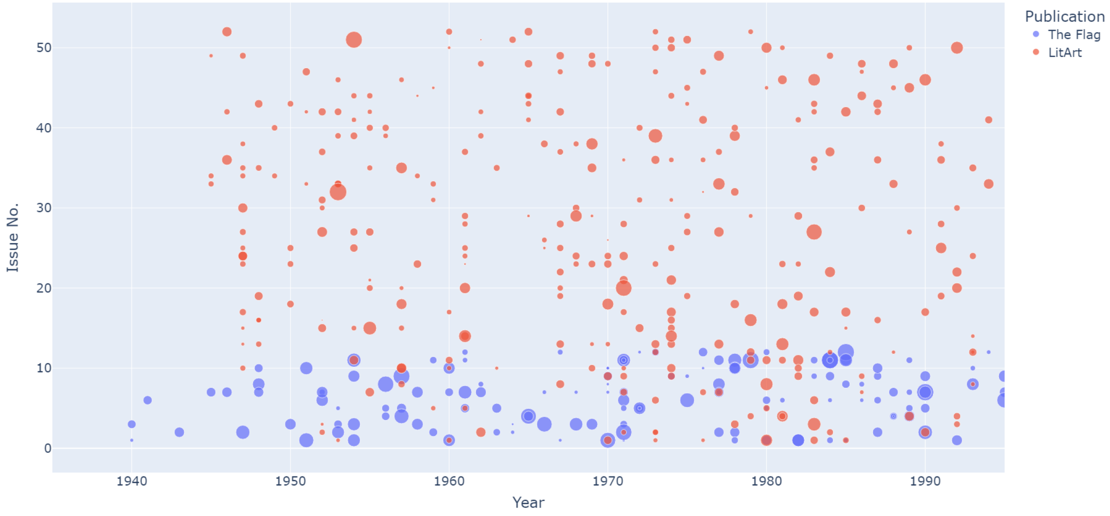


Figure 5: Most popular matching n-grams: "it is self-evident that"

Table 4
Number of similar documents per quantiles

TF-IDF score	0.44 - 0.49	0.49 - 0.50	0.50 - 0.52	0.52 - 0.53	0.53 - 0.68
Seed 1	37	9	1	2	
Seed 2		8	16	11	14
Seed 3	12	20	7	7	3
Seed 4		12	17	11	9
Seed 5			8	18	23
Doc2Vec score	0.64 - 0.66	0.66 - 0.67	0.67 - 0.69	0.69 - 0.71	0.71 - 0.78
Seed 1	15	11	7	9	7
Seed 2		14	11	13	11
Seed 3	12	12	13	6	6
Seed 4	22	12	5	6	4
Seed 5			13	15	21

4 shows the number of similar documents, divided into quintiles (i.e., groupings of documents in five sets from lower to higher similarity levels). Although the scores of TF-IDF and Doc2Vec embeddings are not entirely congruent, there is a tendency for results to overlap.

For faster visual exploration all three similarity metric scores were uploaded into Neo4J graph database as edges connecting vertices - documents. Figure 6 shows most similar documents to all 5 seeds, including all three types of embeddings (5-gram, TF-IDF, Doc2Vec) in the network visualization. The results are filtered with the cutoff being the top quintile for the corresponding metric. The Neo4J visualization tool allows the researcher to browse the relationships and inspect similar articles interactively - Figure 6 is a snapshot from this environment.

The qualitative analysis of the similar articles reveals that in the majority of cases, the most similar documents are discussing the same topics as the seed articles. Regarding the distribution of similar articles across time, there were differences among the seeds. For instance, the majority of similar documents related to the Seed 5 ("Larks call to battle") were published in 1940s, i.e., close to the time of the publication of the seed article, followed by several publications in 1950s and very few in the later years. It can be hypothesized that this correlates with the dynamics of the popularity and relevance of poet Jānis Sudrabkalns and this poetry collection in particular.

There are fewer similar documents in the highest quintile for Seeds 1, 2, 3, and 4, compared to Seed 5, and the articles are published at different times from the 1940s to 1 990s. Seeds 3 and 4, which are both discussing visual arts, share the largest number of similar documents. Documents similar to 3 and 4 are historical and theoretical accounts of the situation of visual art in Latvia, written in the style resembling Seed 3, which is not as much propagandist but rather a scholarly article devoted to the theory of Socialist Realism in art. Seed 2, the "Declaration of the Writers of Soviet Latvia" is surrounded, first, by documents published close in time to the original document and, second, by documents issued around the time of celebrating anniversaries of the founding of the Union of Latvian Soviet Writers: to a large extent, this last set of documents overlaps with the articles that were identified as containing direct citations in the analysis of matching 5-grams.

Overall, a close inspection of the relations of each individual seed article reveals that the document similarity networks, first of all, tell stories about the situational and historical relevance of particular subjects discussed in these documents - publication of a fashionable poetry book, the founding of the Latvian Soviet Writers' Union, the opening of the first Latvian Soviet art exhibition. It is especially pronounced amidst the most similar articles that belong to quantile with the highest similarity scores.

When casting the net more widely and looking at documents with slightly lower similarity scores, there is more variation, however, it is difficult to judge to what extent the ideologically-driven stylistic and rhetorical devices are contributing to the similarity scores - perhaps, the subject matter of the document always plays the leading role. One can argue that the distinction between the subject matter and rhetorical devices is not that important here since both aspects are inseparably intertwined in the vocabularies of discourses and a particular Zeitgeist dictates simultaneously the topics and the rhetoric. A further meta-analysis of the similarity networks of all articles would be needed to find out whether there are global patterns of document similarity that could be at least in part attributed to the discourse change.

5. Conclusions

It was found that identifying matching 5-grams reliably works for detecting direct citations. Five and more matching 5-grams indicated direct citation (2 to 8 citations for each seed article).

5-gram matches with the length of 1 to 4 consecutive matching 5-grams often entail frequently used phrases: (1) ideological aphorisms directly related to SR doctrine ("nationalist in form, socialist in content"); (2) other idioms ("nation's best sons and daughters"); (3) named entities ("communist party", "soviet union"); (4) multi-word expressions ("it is self-evident that").

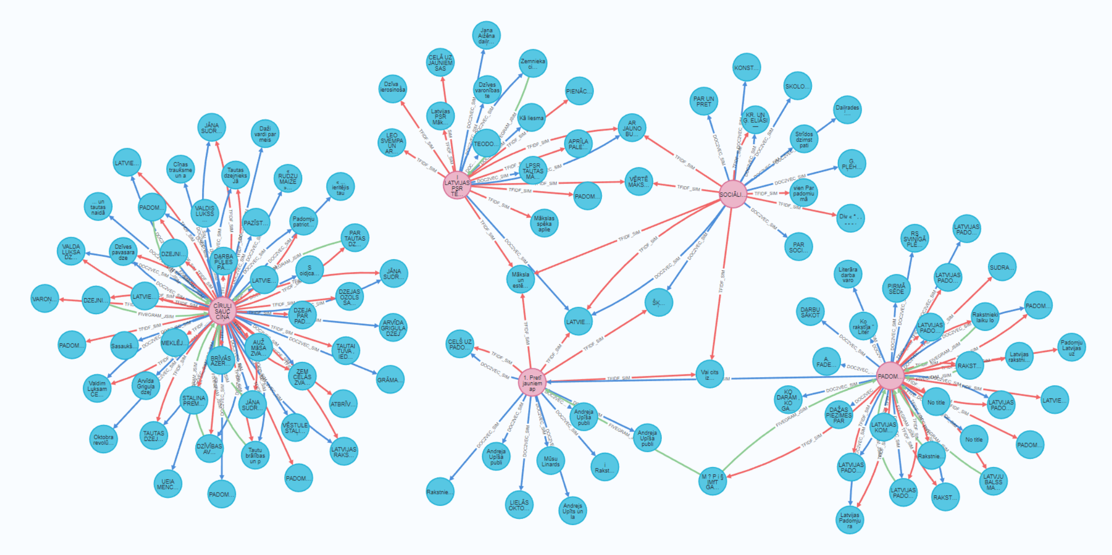


Figure 6: Documents similar to Seed documents by top quintile scores TF-IDF (>0.53) or Doc2Vec (>0.71) (Neo4j visualization)

The TF-IDF and Doc2Vec results for documents most similar to the seed documents showed almost 90 percent overlap despite different approaches: tf-idf calculation performed on the word level, Doc2Vec on the phrase level. Articles selected as most similar to the seed documents: (1) are on a similar topic; (2) contain similar ideological vocabulary; (3) were published closer to the date of the publication of the seed article in several cases.

In the current form, TF-IDF and Doc2Vec-based document similarity results are suitable for exploratory analysis, further research would be required to obtain hard evidence on the proliferation of the discourse in time and across several newspapers. Compared to the methodology aimed at finding similar documents, analysis of the reuse of individual phrases and passages across time (i.e., analysis of matching n-grams) can provide faster results that are immediately usable for qualitative research.

The authors hypothesize that both approaches selected for the case study are very promising for further usage in discourse studies. The proliferation of the discourse can be traced by identifying direct citations of notable works and dissemination of individual eminent phrases, as well as by analyzing the similarity of documents. More quantitative and qualitative analysis is required to provide evidence-based generalizations on the overall proliferation of the Socialist Realism discourse in the given Corpus. Further studies could be based either on the analysis of a considerably larger number of seed articles carefully selected from the same corpus or other sources or performed as a meta-analysis of the similarity networks of all articles in the Corpus.

Acknowledgments

The study was conducted with the support of the project "Text Analysis Methods and Tools for Similarity Metrics in Large National Text Corpora: the Case of the Latvian National Digital Library and the National Repository of Academic Texts of Ukraine", Latvian-Ukrainian bilateral cooperation programme, funded by the Latvian Council of Science and State Education Development Agency https://lnb.lv/en/about-us/projekti/.

References

[1] T. Foltýnek, N. Meuschke, B. Gipp, Academic plagiarism detection: A systematic literature review, ACM Comput. Surv. 52 (2019). URL: https://doi.org/10.1145/3345317. doi:10.1145/3345317.

[2] H. Salmi, P. Paju, H. Rantala, A. Nivala, A. Vesanto, F. Ginter, The reuse of texts in finnish newspapers and journals, 1771–1920: A digital humanities perspective, Historical Methods: A Journal of Quantitative and Interdisciplinary History 54 (2021) 14–28. URL: https://doi.org/10.1080/01615440.2020.1803166. doi:10.1080/01615440.2020.1803166. arXiv:https://doi.org/10.1080/01615440.2020.1803166.

[3] D. A. Smith, R. Cordell, A. Mullen, Computational methods for uncovering reprinted texts in antebellum newspapers, American Literary History 27 (2015) E1–E15.

[4] W. H. Gomaa, A. A. Fahmy, et al., A survey of text similarity approaches, international journal of Computer Applications 68 (2013) 13–18.

[5] G. Jēkabsons, Evaluation of fingerprint selection algorithms for local text reuse detection, Applied Computer Systems 25 (2020) 11–18.

[6] M. Zachara, D. Pałka, Comparison of text-similarity metrics for the purpose of identifying identical web pages during automated web application testing, in: Information systems architecture and technology: Proceedings of 36th international conference on information systems architecture and technology–ISAT 2015–Part II, Springer, 2016, pp. 25–35.

[7] J. P. Gee, M. Handford, Introduction (pp. 1-6), 2012.

[8] N. Fairclough, Critical discourse analysis, in: The Routledge handbook of discourse analysis, Routledge, 2013, pp. 9–20.

[9] E. Britannica, Socialist realism, 2022. URL: https://www.britannica.com/art/Socialist-Realism, accessed March 16th, 2023.

[10] P. Zeile, Sociālistiskais reālisms, Liesma (1981). Rīga.

[11] H. Günther, How socialist realism was exported to eastern european countries and how they got rid of it, in: Socialist Realism in Central and Eastern European Literatures Under Stalin, Anthem Press, 2018, p. 17. Eds E. Dobrenko and N. Jonsson-Skradol.

[12] E. Dobrenko, N. Jonsson-Skradol, Socialist Realism in Central and Eastern European Literatures Under Stalin, volume 1, Anthem Press, 2018.

[13] S. Pelše, Sociālistiskais reālisms laikrakstā "literatūra un māksla". padomju perioda mākslas kritikas teorētiskie pamati, Mākslas Vēsture un Teorija 1 (2003) 16–23. [1691-0869].

[14] S. Schleimer, D. S. Wilkerson, A. Aiken, Winnowing, in: Proceedings of the 2003 ACM SIGMOD international conference on Management of data, ACM, 2003. URL: https://doi.org/10.1145/872757.872770. doi:10.1145/872757.872770.

[15] G. Bonaccorso, Machine Learning Algorithms: Popular algorithms for data science and machine learning, Packt Publishing Ltd, 2018.

[16] Q. Le, T. Mikolov, Distributed representations of sentences and documents, in: International conference on machine learning, PMLR, 2014, pp. 1188–1196.

[17] J. Han, M. Kamber, J. Pei, Cosine similarity, in: Data Mining: Concepts and Techniques, Elsevier, Morgan Kauffman, 2012, pp. 74–76.

[18] A. Znotiņš, E. Cīrule, Nlp-pipe: Latvian nlp tool pipeline, Human Language Technologies–The Baltic Perspective (2018) 183–189.

Finding environmental discourse in historical newspapers: a topic model workflow for query disambiguation

Peeter Tinits[1]

[1]University of Tartu, Ülikooli 18, 50090 Tartu, Estonia

Abstract

Digitized historical newspapers are a treasure trove of information for our understanding of the past. As one popular application, the frequencies of query matches can be used to understand the prevalence of some discourse in a historical era. This requires the construction good queries: broad enough to capture diverse contexts and narrow enough to exclude irrelevant ones. For historical research in digital humanities, targeted queries that emphasize precision have been advised. In this paper, we develop an alternative approach, by using broad queries to cast a wider net and then using topic models built on the match contexts to filter out irrelevant matches. Specifically, we look for contexts discussing environmental issues throughout the 20th century using a corpus of two Australian newspapers. We report on a comparison of iteratively constructed narrow and broad queries and their precision and recall, and find our approach to discover roughly 7-10x more matches with a comparable level of accuracy. This combined approach can work well for focussed research projects where deliberate query construction and qualitative feedback on the results is feasible.

Keywords

text mining, keyword frequencies, query disambiguation, topic model, historical inference

1. Introduction

Full text access to digitized historical newspaper collections opens up a range of new research opportunities for humanities and social sciences. A central way to interact with digital archives is via keyword search [1, 2] that is also a built-in feature for most archival collections [3]. The frequencies of keyword matches as a proportion of a representative text corpus have been used to study the prevalence of discourse tied to specific events over time [4, 5] or track long term trends in discourses [6, 7].

A key difficulty for a researcher here is to construct useful keyword sets that are precise enough to exclude most irrelevant matches, but broad enough to include most relevant matches [8, 9]. Historical materials and long time periods exacerbate these concerns: the queries need to rely on historical background knowledge to cover changes in preferred terminology to denote a topic, overcome fluctuations in meanings and alternative use contexts of these terms, and be relatively robust to varying OCR quality (e.g. [8, 9]). Constructing useful queries thus requires a lot of domain expertise and experimentation in finding the right search term for the corpus at hand. Approaches to query construction thus make up a central issue in historical research with large digitized collections [10].

The quality of the query can greatly influence the conclusions drawn from the uncovered keyword frequencies. If the query retrieves also many documents that are irrelevant, i.e. if it has low precision, the frequencies can be overestimated for some periods. If the query retrieves mainly or only documents that are relevant but misses many more relevant documents, i.e. if it has low recall, the frequencies can be underestimated for some periods. In both cases if the query works unevenly well across the time period in question, the estimation of the prevalence of a topic based on keyword frequencies may skew also the historian's interpretation. In information retrieval, a trade-off can be seen in precision and recall, thus forcing a good query to balance between the two [11].

In this paper, we aim to retrieve a diverse set of contexts talking about nature and environmental issues from historical text sources covering almost the whole 20th century. This has grown out of a practical research interest articulated in the Deep Transitions theory [12, 13]: the theory postulates that a major shift in public attention on the natural environment would have taken place from 1960s onwards from a relatively marginal position to a common topic in public discourse. We can study this proposal systematically via keyword frequencies in a representative corpus of texts. However to understand the historical dynamics, we need to consider the broad set of contexts that these discussions took place in, while at the same time maintaining a sensible level of precision in our query responses.

DHNB2023 | Sustainability: Environment - Community - Data. The 7th Digital Humanities in the Nordic and Baltic Countries Conference. Oslo – Stavanger – Bergen, Norway. March 8–10, 2023.

EMAIL: peeter.tinits@ut.ee (P. Tinits)
URL: https://peetertinits.github.io/ (P. Tinits)
ORCID: 0000-0003-3186-6264 (P. Tinits)

DHNB Publications, DHNB2023 Conference Proceedings,
https://journals.uio.no/dhnbpub/issue/view/875

1.1. Constructing good queries

A good query should thus find a way to avoid irrelevant topics, but also allow for some flexibility in the contents [10]. For example, if we want to find discourse on the natural environment from the corpus, looking for the words *nature* or *environment* does not form a very precise query. The word *nature* is used in very different meanings in newspapers - while it often denotes the natural environment we have in mind, it is also frequently used to point to the 'nature of things' (e.g. *human nature*, *the delicate nature of the correspondence*). Same goes for 'environment' - it has been used frequently for any surrounding context ('e.g. *political environment*, *school environment*). These frequent matches form **false positives** for our query, as it was not something we were interested in. Phrases like *nature lovers* or *protection of nature* would on the other likely find consistently only relevant texts, forming **true positives**.

A PRECISE query then aims to cover a sensible amount of relevant terms and phrases. Capturing all is usually not possible, finding the right terms will require good domain expertise and experimentation with the corpus. If we use within a query multiple alternative terms, possibly from different domains (e.g. an ideological term AND a party associated name [14]), we are able to get also texts from different contexts. Thus, we could look for *nature lovers*, *protection of nature*, *ecology*, and *sustainable development* to get mostly relevant texts on different aspects of discourse about nature. An open issue here is that we may miss a crucial part of discourse that would influence also the conclusions about the general trends.

A BROAD query can be easier to construct. Although domain expertise is still needed, it is less dependent on the phraseology within the corpus. On the other hand the low precision can give us a large number of irrelevant texts - enough to interfere with the inferences we would like to make. The words *nature* or *environment* can make good queries of this type. Also for BROAD queries, multiple keywords would be good to cover different times and contexts. If we use such general keywords, we need to find some way to filter out irrelevant queries after the initial search.

In both cases, the query needs to be constructed iteratively, and the sources need to be explored for candidate keywords. We may discover terms or phrases that were very good indicators at an earlier time, alternative meanings to the phrases we used, or systematic errors (e.g. common OCR mistakes) within the texts. This process of iteration is not often systematically described and evaluated in historical research. Here, we describe the usefulness of the queries by exploring their precision and recall through a constructed limited ground truth. We can construct this limited ground truth on the basis of a limited set of terms and phrases that we know to refer to events or matters unambiguously connected with these topics having a very high precision. At the same time, the limited breadth of these keywords makes them poor candidates to study long-term change though queries since those special circumstances present only a fraction of relevant contexts.

1.2. Disambiguating BROAD queries

In information retrieval, a number of techniques have been developed to improve the query results to match a user's intent. Usually, this is made with a general user in mind, so it would work with any query, in any corpus. Commonly, this is done by methods for query expansion (e.g. [15, 16]), query disambiguation [17, 18, 19] or word sense disambiguation [20, 21]. In refining search results on specific queries, topic models in particular have been found to be a useful tool to improve the accuracy of the query [22, 23, 24].

Here, we develop a simple practical workflow that uses BROAD queries to find matches in diverse contexts and annotated topic models to disambiguate between relevant and irrelevant matches in a data-driven way. This approach relies less on historical domain expertise and experimentation with the corpus and allows for easier translation between cultural contexts and languages (e.g. when aiming to compare trends in different countries). We contrast this approach with the traditional PRECISE queries.

2. Data and methods

2.1. Data

The selection of data in studying 20th century trends is limited by data availability and access [2]. Here, we focus on Australian materials which have been combined to cover 1900-1995: *Sidney Morning Herald* (1900-1940) and *Canberra Times* (1930-1995). The newspaper texts were acquired from the Trove archives via GLAM workbench tools [25].

We lemmatized the text with the spaCy NLP package for Python (v. 2.1) that has been found to be more robust to OCR noise than alternatives [26]. We excluded pronouns and determiners from the analysis. Altogether there were 4,874,397 articles in the set constituting 749,867,456 tokens after this preprocessing. See Figure 1 on articles and tokens per year.

2.2. PRECISE and BROAD queries

Following the guidelines of specialist historians in the group, two sets of keywords were constructed to find examples of popular discourse on nature, with a focus on nature protection and conservation: PRECISE and BROAD.

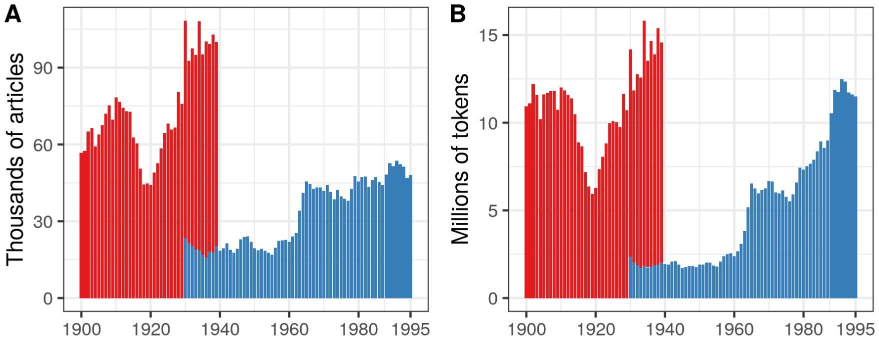


Figure 1: The sources used in the example. A: Thousands of articles per year, B: Millions of tokens per year. Sidney Morning Herald (1900-1940) in red, Canberra Times (1930-1995) in blue.

Table 1

The iterative construction of the keyword queries. P - precise, B - broad. S marks the seed terms, IT1-4 point to subsequent iterations. X denotes a natural object represented as forest*/animal*/bird*/tree*/water*/ wildlife*/nature*/soil*/land*. Asterisk (*) in the table indicates any additional characters. For all searches, case was ignored, word boundaries were used as limits.

name	type	contents
P_S	seed	nature protection/conservation; environmental protection/regulation; environmentalis*; conservationis*; ecolog*; sustainab*; biodivers*
P_{IT1}	add	conservation of (the) X; protection of (the) X
P_{IT2}	add	X conservation; X protection;
B_S	seed	natur[ea]*; environment*
B_{IT1}	add	conservation*
B_{IT2}	add	sustainab*
B_{IT3}	add	earth
B_{IT4}	add	naturalist; ecolog*; pollution

The queries were then iteratively expanded through experimention with the corpus and comparing the findings to specialist reference literature. See Table 1 on the exact terms used.

For the PRECISE set, a number of general phrases that should be unambiguously related to this topic were collected. Reviewing the results and the related literature (e.g. [27]), it was found that *conservation* was a significant keyword in earlier periods and added phrases with this in the first iteration. Based on the experiments with the corpus, we also added alternative phrasings to the seed words in a second iteration that proved very common as well.

For the BROAD set, the keywords *nature* and *environment* were used as the initial seed set. Similarly, *conservation* was added as a significant term in the first iteration. Then a few more general terms were added to capture the diversity of topics present based on further experimentation with the corpus.

2.3. Ground truth

We constructed a set of ground truth articles (n = 9070) based on another set of queries, on known terms and events that are definitely relevant, but limited in scope. If our queries work well, we would expect them to retrieve also the articles in this constructed limited ground truth. The keyword set was composed based on the suggestions of three specialist historians on the team and the historical sources on the history of environmentalism in Australia and around the world (e.g. [27]). See Table 2 for details.

2.4. Disambiguating BROAD query results via topic models

We constructed topic models on the basis of the search results for each query. For this: 1) We extracted snippets of +/-25 words next to query response. 2) We filtered the words into 4 different input sets: unfiltered raw snippets in lowercase; lemmatized snippets in lowercase; common lemmas ($n > 10$ in corpus) in lowercase; frequent

Table 2
Ground truth keywords. X denotes a natural object represented as forest*/animal*/bird*/ tree*/water*/ wildlife*/nature*/soil*/land*

Type	terms
Known global issues	acid rain; greenhouse effect; deforestation; overfishing; ocean acidification; soil degradation; soil erosion; desertification; dust bowl
Professional keywords	ecosystem; biodivers*
Specific names/terms	gaia; rachel carson
Local events/issues	wildlife preservation*; koala killing*
Institutional terms	nature protection; nature conservation
Common phrases	conservation of (the) X; protection of (the) X; X conservation; X protection

Table 3
Recall and precision for queries by type for selected queries. * - For precise queries, maximal precision is assumed. F-score denotes the harmonic mean of precision and recall.

		With filter	Matches (keywords)	Matches (articles)	Recall per keyword	Recall per text	Precision	F-score
Precise	S		11640	10074	0.27	0.15	1^*	0.26
Precise	IT2		17671	15595	0.51	0.76	1^*	0.86
Broad	S	-	379545	270668	0.57	0.39	0.21	0.27
Broad	**S**	+	**114532**	**73996**	**0.51**	**0.32**	**0.83**	**0.46**
Broad	IT3	-	491057	323921	0.74	0.86	0.30	0.44
Broad	**IT3**	+	**192754**	**119248**	**0.68**	**0.78**	**0.67**	**0.72**
Broad	IT4	-	510851	329523	0.75	0.86	0.31	0.46
Broad	**IT4**	+	**184656**	**104320**	**0.68**	**0.76**	**0.77**	**0.76**

lemmas ($n > 100$ in corpus) in lowercase. 3) We built up topic models with the common LDA [28] algorithm using MALLET ([29, 30]). For each set in (2), we built models of 10, 20, and 30 topics. 4) Within each of the topic models for each set (12 models per query), we annotated the topic models for relevance. 5) We considered a match a true positive only if the snippet contained at least 20% relevant topics in at least 8/12 models. Here we tried several parameters, see Table A1 for details.

2.5. Measuring precision and recall

We measure recall as a simple metric of the percentage of ground truth articles that include a match (see 2.3). We do this across texts, and separately for each keyword in the ground truth. We report the average across keywords then. Here, we simplify variations of *protection of**, *conservation of **, ** conservation*, and ** protection* into general types.

We measure precision by taking a random sample of 100 snippets that have successfully passed the topic model filter and mark each snippet as either a true positive or a false positive based on the relevance to the query.

3. Results

3.1. Influence on precision and recall

When we compare the two approaches (see Table 3) we find that BROAD queries expectedly find many more matches even after the application of the topic model filter 20-25x more with no filter, 7-10x more with filter). The recall of the ground truth texts is comparable and sometimes better for BROAD queries. While the precision of BROAD queries is low without a filter (ca. 0.20-0.30), applying a filter greatly improves this, reaching ca. 0.70-0.80 precision which may be considered a tolerable degree of error. There is a degree of randomness involved, and with higher filter thresholds, the model can find even 0.95 precision (see Table A1) though at the cost of recall. Of the iterations, recall greatly increased by Iteration 2 of the PRECISE queries, while Iterations 2-4 are rather similar in outcomes for BROAD queries. We pick Iteration 3 as the one best option here, as it had the highest recall among them, and a simpler query than Iteration 4. These choices of which threshold to pick and which query to select naturally depend on the questions at hand.

Figure 2 presents the recall of the ground truth visu-

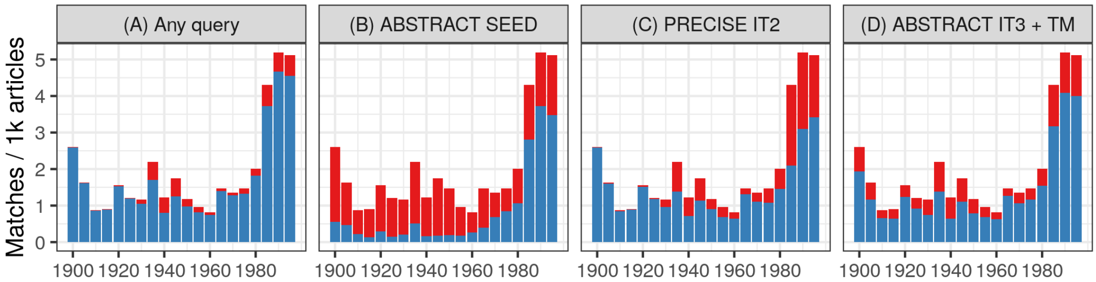


Figure 2: The recall of different approaches by five year periods as a proportion of articles. A: Articles found by any query. B: Initial seed for broad terms. C: Articles found by best precise query (Precise, iteration 2). D: Articles found by best query via topic models (Broad, iteration 3). The height shows the number of matches per 1,000 articles in each 5-year slice. Blue matches were successfully found, red matches were not.

ally. The ground truth found 9070 articles distributed across the period, with a few more at the end. Most of the articles were retrieved by at least one query. The best PRECISE query retrieved most of the texts across the period, but missed comparatively more texts at the very end. The BROAD seed query behaved rather poorly (0.39 recall) across the period due to the lack of 'conservation' in search. The best BROAD query with a filter retrieved most of the articles, and covered the period more or less equally, thus giving a more balanced result than the PRECISE query.

3.2. Impact on historical inference

Figure 3 presents the prevalence of discourse by different measures. We look at it from three different visualization perspectives, as they are often used in keyword frequency studies. Figure 3-A shows the trends on linear scale - a simple way to show absolute changes; Figure 3-B shows trends on a logarithmic scale - useful when the magnitude of the changes is in question (e.g. when small quantities grow 10x); Figure 3-C shows the trends on a relative scale - linear growth compared to the maximal position reached. When analysing changes in the level of discourse, looking at relative trends can be an important tool: while the set of keywords may not capture all the texts, they can still show increases and decreases in discourse, assuming mostly even coverage over time.

Visibly, the broad queries found many more texts, with and without the filter. Without the filter, the relative prevalence of texts is higher 1900-1940, which is then likely due to the false positives. With the filter, the trends in the BROAD queries are comparable to the best PRECISE query (see Figure 3-C). In relative terms, the PRECISE query estimates there to be less discourse from 1970s to mid 1980s than the BROAD query. Different iterations of the queries are very similar for broad terms, while the first two iterations of PRECISE queries seem to underestimate the number of matches before the 1970s.

Figure 4 presents the comparative results of one of these models from each type - the "best one" (P-IT2, B-IT3-NOFILTER, B-IT3). Each of the models shows a growth in frequency starting from around 1950s, although the BROAD query results without a filter estimate (falsely) the discussions to be much more prevalent in early years (due to many irrelevant matches for broad terms). The growth starts at a similar place for each, estimated by the first derivative to be around 1950-1955. The relative growth is very similar between the PRECISE query and the BROAD query with the filter, although the estimate by the PRECISE query is comparably lower from 1970s to mid 1980s. We know from the distribution of the missed texts (see Figure 2) that the precise query is somewhat biased against the quick growth of relevant articles found in ground truth set starting from the 1980s. For this reason, we expect that between the two strategies, the BROAD query with the filter is more accurate within the relative measure.

4. Discussion

The results do conform with the expectations of the Deep Transitions theory, a significant change in the prevalence of discourse on natural environment can indeed be seen from around the 1960s. 1960s gave rise to popular environmental movements that brought the issues of human impact on natural environment into focus (see e.g. [27]). The study on keyword frequencies can show the breadth of contexts where these discussions appeared. Similar results can be found both with the PRECISE query and BROAD query, if the topic model filter is applied. However, when the study aims to build on the extracted matches (e.g. in analysing the contexts of the matches),

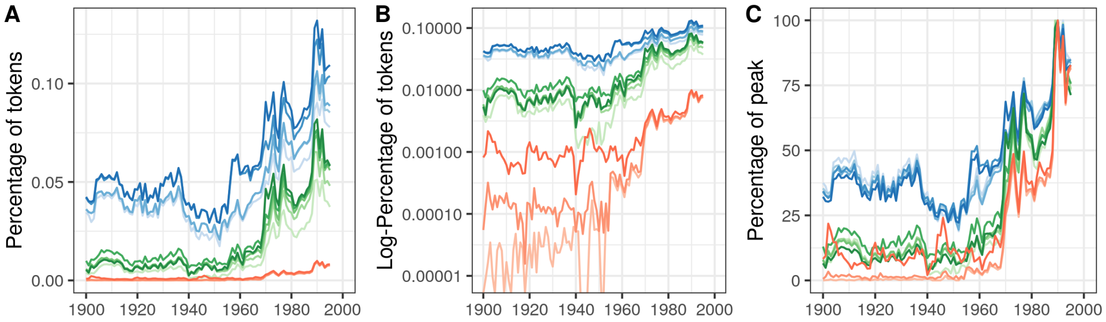


Figure 3: Results of the queries colored by type and iteration. Blue - broad query, no filter. Green - broad query - with filter. Red - precise query. The lightest line is the seed set of keywords, colors get darker with each iteration. A: Frequencies on a linear scale; B: Frequencies on a logarithmic scale, C: Frequencies as a proportion of their maximal value.

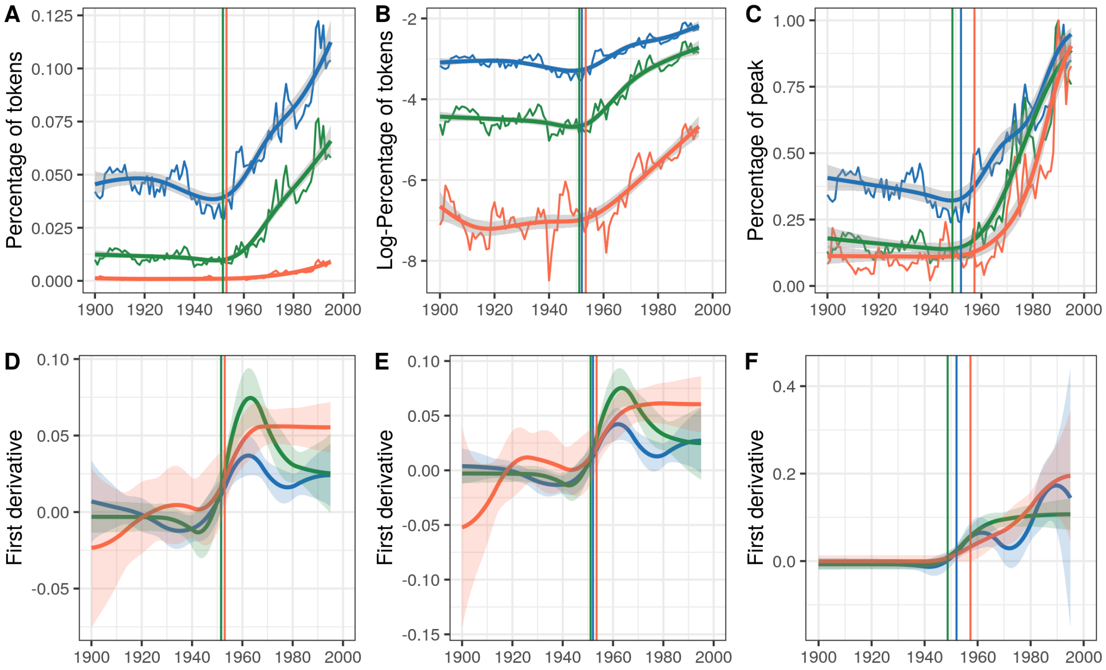


Figure 4: Search results of the best model for historical interpretation.A: Frequencies on a linear scale; B: Frequencies on a logarithmic scale, C: Frequencies as a proportion of their maximal value. D-F: The first derivatives on the graphs above them. The vertical lines in each graph show the beginning of the period of growth - the point when the first derivative was significantly different from 0.

the BROAD query can capture a much bigger density of contexts with reasonable precision. This can then be used for follow-up analysis, such as analysing the sentiments within each found match (as done in [31]). The value of different approaches to keyword frequencies always depend on the constraints and possibilities of a particular research project.

Constructing useful queries for historical research necessarily needs to balance multiple aspects: ease of use, data quality, need for expertise, and background historical changes. PRECISE queries may be computationally easy to perform, but require a lot of specialized competence and experimentation with the corpus. BROAD queries require an extra step of disambiguation in a computational workflow, but common LDA topic models are capable of doing the job. There are a number of different algorithms and implementations to use topic models to summarise texts with specific use cases and limitations. We opted to use LDA as it is the most common simple application of this, however it's conditions of use have been subject to criticism and its capacities have been greatly expanded in newer models [32]. Further improvements are expected from using dynamic topic models [33, 34], structural topic models [35, 36] or pseudo-document based topic models [37] for the filters. The principle of using topic models to find relevant query matches can remain also in this case.

The results presented here - in estimating the prevalence of discourse on environmental issues over the 20th century - remain dependent on qualitative choices made during the study. The choice of initial seed keywords, the experiments performed with the corpus, and the methods used to improve the query results here with topic models may have been made differently and led to a different path. The assessment of the approach through a limited constructed ground truth and the experiments done with different parameters of the algorithms increase researcher confidence in the results, however they can not be said to be the only possible interpretation. The use of query construction and its algorithmic refinement in historical research will thus rely on interdisciplinary collaboration with both historians and data scientists playing an important role. Precise guidelines can thus not be given here.

The precise parameters to be set in filtering (e.g. which threshold to use, how many topic models to annotate and how big, does it work with one large topic) need to be experimented with, and the results may depend on the query, topic, or the corpus in focus. Here, 20% from 8/12 models, or 33% from 6/12 models worked well, so for similar corpora and questions, they may do so too.

Principled workflows for query construction, while emphasised by historians as crucial [1, 10] are still somewhat underdeveloped. We suggest that the use of selected keywords that are knowingly limited in scope to construct a ground truth can be a way to help this prospect. Being able to clearly track the recall of expected true positives is an asset for query construction, and such a ground truth query may be a simple way to do this.

Much of the research in information retrieval is built around a general user to effectively extract any information from a text corpus. Digital humanities researchers can be a special user case, where it is feasible to suggest longer workflows involving qualitative feedback from the researcher to improve the accuracy of the query. The use of topic models to filter out irrelevant contexts from deliberately constructed queries can offer one solution in this niche. It improves the potential for historical inference from keyword queries on a representative corpus to assess prevalence of particular discourse over time. This can be improved or adjusted by the use of other techniques in word sense disambiguation or query disambiguation. In general, a systematic approach to query construction based on theories from information retrieval can be a benefit to the digital humanities researchers, however the implemented solutions can aim to get the best of both worlds: designed custom workflows based around text processing tools as well as qualitative feedback and insight from a researcher based on domain expertise. This opens up also room for more specialized methods that focus on limited use cases rather than general ways to find relevant information.

Acknowledgments

This work was supported by the Estonian Research Council grant PRG346 "Reshaping Estonian energy, mobility and telecommunications systems on the verge of the Second Deep Transition". Important contributions were made by Laur Kanger, Olaf Mertelsmann, Erki Tammiksaar and Aro Velmet in setting the ground for the study, helping with historical expertise and and keyword selection; and Stefan Hartmann, Simon Hengchen, Andres Karjus, Artjoms Šeļa, and Oleg Sobchuk in helping with the technical workflow and providing very useful feedback on versions of this paper.

References

[1] B. Nicholson, THE DIGITAL TURN: Exploring the methodological possibilities of digital newspaper archives, Media History 19 (2013) 59–73. doi:10.1080/13688804.2012.752963.

[2] T. Smits, Problems and possibilities of digital newspaper and periodical archives, Tijdschrift voor Tijdschriftstudies 0 (2014) 139. doi:10.18352/ts.317.

[3] M. Ehrmann, E. Bunout, M. Düring, Historical Newspaper User Interfaces: A Review, in: IFLA WLIC 2019 - Athens, Greece - Libraries: Dialogue for Change, 2019.

[4] T. Lansdall-Welfare, S. Sudhahar, J. Thompson, J. Lewis, FindMyPast Newspaper Team, N. Cristianini, Content analysis of 150 years of British periodicals, Proceedings of the National Academy of Sciences 114 (2017) E457–E465. doi:10.1073/pnas.1606380114.

[5] R. J. Shiller, Narrative Economics: How Stories Go Viral and Drive Major Economic Events, Princeton University Press, Princeton, 2019.

[6] J.-B. Michel, Y. K. Shen, A. P. Aiden, A. Veres, M. K. Gray, The Google Books Team, J. P. Pick-

ett, D. Hoiberg, D. Clancy, P. Norvig, J. Orwant, S. Pinker, M. A. Nowak, E. L. Aiden, Quantitative Analysis of Culture Using Millions of Digitized Books, Science 331 (2011) 176–182. doi:10.1126/science.1199644.

[7] Q. Van Galen, B. Nicholson, In Search of America: Topic modelling nineteenth-century newspaper archives, Digital Journalism 6 (2018) 1165–1185. doi:10.1080/21670811.2018.1512879.

[8] B. Nicholson, Counting Culture; or, How to Read Victorian Newspapers from a Distance, Journal of Victorian Culture 17 (2012) 238–246. doi:10.1080/13555502.2012.683331.

[9] P. M. Greenfield, The Changing Psychology of Culture From 1800 Through 2000, Psychological Science 24 (2013) 1722–1731. doi:10.1177/0956797613479387.

[10] H. Huistra, B. Mellink, Phrasing history: Selecting sources in digital repositories, Historical Methods: A Journal of Quantitative and Interdisciplinary History 49 (2016) 220–229. doi:10.1080/01615440.2016.1205964.

[11] C. D. Manning, P. Raghavan, H. Schütze, Introduction to Information Retrieval, Cambridge University Press, Cambridge, UK, 2008. URL: http://nlp.stanford.edu/IR-book/information-retrieval-book.html.

[12] J. Schot, L. Kanger, Deep transitions: Emergence, acceleration, stabilization and directionality, Research Policy 47 (2018) 1045–1059. doi:10.1016/j.respol.2018.03.009.

[13] L. Kanger, J. Schot, Deep transitions: Theorizing the long-term patterns of socio-technical change, Environmental Innovation and Societal Transitions 32 (2019) 7–21. doi:10.1016/j.eist.2018.07.006.

[14] H. Piersma, I. Tames, L. Buitinck, J. van Doornik, m. marx, War in parliament: What a digital approach can add to the study of parliamentary history, Digital Humanities Quarterly 8 (2014).

[15] B. Selvaretnam, M. Belkhatir, Natural language technology and query expansion: Issues, state-of-the-art and perspectives, Journal of Intelligent Information Systems 38 (2012) 709–740. doi:10.1007/s10844-011-0174-3.

[16] H. K. Azad, A. Deepak, Query expansion techniques for information retrieval: A survey, Information Processing and Management 56 (2019) 1698–1735. URL: https://www.sciencedirect.com/science/article/pii/S0306457318305466. doi:https://doi.org/10.1016/j.ipm.2019.05.009.

[17] F. Sadat, Research on Query Disambiguation and Expansion for Cross-Language Information Retrieval, Communications of the IBIMA (2010) 1–11. doi:10.5171/2010.438404.

[18] H. Zhang, K. Yang, E. Jacob, Topic Level Disambiguation for Weak Queries, arXiv:1502.04823 [cs] (2015). doi:10.1633/JISTaP.2013.1.3.3. arXiv:1502.04823.

[19] P. Kotoula, C. Makris, Query disambiguation based on clustering techniques, in: L. Iliadis, I. Maglogiannis, V. Plagianakos (Eds.), Artificial Intelligence Applications and Innovations, Springer International Publishing, Cham, 2018, pp. 133–145.

[20] J. Boyd-Graber, D. Blei, X. Zhu, A topic model for word sense disambiguation, in: Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL), Association for Computational Linguistics, Prague, Czech Republic, 2007, pp. 1024–1033. URL: https://aclanthology.org/D07-1109.

[21] L. Li, B. Roth, C. Sporleder, Topic models for word sense disambiguation and token-based idiom detection, in: Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Uppsala, Sweden, 2010, pp. 1138–1147. URL: https://aclanthology.org/P10-1116.

[22] X. Wei, W. B. Croft, Lda-based document models for ad-hoc retrieval, in: Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '06, Association for Computing Machinery, New York, NY, USA, 2006, p. 178–185. URL: https://doi.org/10.1145/1148170.1148204. doi:10.1145/1148170.1148204.

[23] M. Erlin, Topic modeling, epistemology, and the english and german novel, Journal of Cultural Analytics 2 (2017). doi:10.22148/16.014.

[24] S. Oberbichler, E. Pfanzelter, Topic-specific corpus building: A step towards a representative newspaper corpus on the topic of return migration using text mining methods, Journal of Digital History 1 (2021).

[25] T. Sherratt, GLAM-Workbench/trove-newspaper-harvester, Zenodo, 2019. doi:10.5281/ZENODO.3545045.

[26] M. Ehrmann, M. Romanello, S. Clematide, P. B. Ströbel, R. Barman, Language resources for historical newspapers: The impresso collection, in: Proceedings of the 12th Language Resources and Evaluation Conference, European Language Resources Association, Marseille, France, 2020, pp. 958–968.

[27] P. Warde, L. Robin, S. Sörlin, The Environment: A History of the Idea, JHU Press, Baltimore, Maryland, 2018.

[28] D. M. Blei, A. Y. Ng, M. I. Jordan, Latent dirichlet allocation, Journal of machine Learning research 3 (2003) 993–1022.

[29] A. K. McCallum, Mallet: A machine learning for language toolkit, http://mallet. cs. umass. edu (2002).

[30] L. Yao, D. Mimno, A. McCallum, Efficient methods for topic model inference on streaming document collections, in: Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2009, pp. 937–946.

[31] L. Kanger, P. Tinits, A.-K. Pahker, K. Orru, A. K. Tiwari, S. Sillak, A. Šeļa, K. Vaik, Deep transitions: Towards a comprehensive framework for mapping major continuities and ruptures in industrial modernity, Global Environmental Change 72 (2022) 102447. URL: https://www.sciencedirect.com/science/article/pii/S0959378021002260. doi:https://doi.org/10.1016/j.gloenvcha.2021.102447.

[32] I. Vayansky, S. A. Kumar, A review of topic modeling methods, Information Systems 94 (2020) 101582.

[33] D. M. Blei, J. D. Lafferty, Dynamic topic models, in: Proceedings of the 23rd International Conference on Machine Learning - ICML '06, ACM Press, Pittsburgh, Pennsylvania, 2006, pp. 113–120. doi:10.1145/1143844.1143859.

[34] J. Marjanen, E. Zosa, S. Hengchen, L. Pivovarova, M. Tolonen, Topic modelling discourse dynamics in historical newspapers, in: S. Reinsone, I. Skadiņa, A. Baklāne, J. Daugavietis (Eds.), Digital Humanities in the Nordic Countries 2020, number 2865 in CEUR Workshop Proceedings, CEUR-WS.org, Germany, 2021, pp. 63–77. URL: http://dig-hum-nord.eu/conferences/dhn2020/, digital Humanities in the Nordic Countries, DHN 2020 ; Conference date: 21-10-2020 Through 23-10-2020.

[35] M. E. Roberts, B. M. Stewart, D. Tingley, E. M. Airoldi, et al., The structural topic model and applied social science, in: Advances in Neural Information Processing Systems Workshop on Topic Models: Computation, Application, and Evaluation, volume 4, 2013.

[36] A. Küsters, E. Garrido, Mining PIGS. A structural topic model analysis of Southern Europe based on the German newspaper *Die Zeit* (1946-2009), Journal of Contemporary European Studies (2020) 1–17. doi:10.1080/14782804.2020.1784112.

[37] Y. Zuo, J. Wu, H. Zhang, H. Lin, F. Wang, K. Xu, H. Xiong, Topic modeling of short texts: A pseudo-document view, in: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, 2016, pp. 2105–2114.

Table A1
Summary of precision and recall of the texts filtered with different parameter thresholds. From iteration one with topic filter.

Set	N	Recall per keyword	Recall per text	Precision
Sum > .2 in 50% models	115992	0.66	0.77	0.66
Sum > .2 in 66% models	94552	0.63	0.73	0.72
Sum > .2 in 75% models	82766	0.61	0.70	
Sum > .33 in 50% models	85886	0.61	0.67	0.86
Sum > .33 in 66% models	68927	0.57	0.61	0.95
Sum > .5 in 50% models	56735	0.51	0.51	
Sum > .5 in 75% models	36351	0.41	0.40	

Table A2
Recall and precision for queries by type. * - For precise queries, maximal precision is assumed. F-score denotes the harmonic mean of precision and recall.

		Filter threshold	Keyword matches	Article matches	Recall per keyword	Recall per text	Precision	F-score
Precise	S		11640	10074	0.27	0.15	1*	0.26
Precise	IT1		12468	10796	0.39	0.23	1*	0.37
Precise	IT2		17671	15595	0.51	0.76	1*	0.86
Broad	S	-	379545	270668	0.57	0.39	0.21	0.27
Broad	**S**	**TH.20**	**114532**	**73996**	**0.51**	**0.32**	**0.83**	**0.46**
Broad	**S**	**TH.33**	**107121**	**68474**	**0.49**	**0.31**	**0.71**	**0.43**
Broad	IT1	-	410954	283265	0.69	0.83	0.32	0.46
Broad	**IT1**	**TH.20**	**155369**	**94552**	**0.63**	**0.73**	**0.72**	**0.72**
Broad	**IT1**	**TH.33**	**143800**	**85886**	**0.61**	**0.67**	**0.86**	**0.75**
Broad	IT2	-	414253	284452	0.7	0.84	0.28	0.42
Broad	**IT2**	**TH.20**	**169864**	**103131**	**0.66**	**0.76**	**0.73**	**0.74**
Broad	**IT2**	**TH.33**	**161660**	**96641**	**0.64**	**0.71**	**0.80**	**0.75**
Broad	IT3	-	491057	323921	0.74	0.86	0.30	0.44
Broad	**IT3**	**TH.20**	**192754**	**119248**	**0.68**	**0.78**	**0.67**	**0.72**
Broad	**IT3**	**TH.33**	**187327**	**114586**	**0.65**	**0.73**	**0.78**	**0.75**
Broad	IT4	-	510851	329523	0.75	0.86	0.31	0.46
Broad	**IT4**	**TH.20**	**184656**	**104320**	**0.68**	**0.76**	**0.77**	**0.76**
Broad	**IT4**	**TH.33**	**172095**	**95401**	**0.65**	**0.70**	**0.75**	**0.72**

A. Appendix

Table A3
Ground truth recall by keyword.

Keyword	N	P$_{SEED}$	P$_{IT2}$	A$_{SEED}$	A$_{SEED+ FILTER}$	A$_{IT2}$	A$_{IT2+ FILTER}$
biodivers*	212	1	1	0.87	0.86	0.95	0.95
conservation of X	490	0.11	1	0.43	0.36	1	0.95
X conservation	5220	0.11	1	0.28	0.22	1	0.92
soil degradation	61	0.33	0.43	0.8	0.77	0.89	0.87
ecosystem	394	0.48	0.51	0.77	0.72	0.86	0.81
deforestation	233	0.35	0.39	0.71	0.62	0.83	0.77
wildlife preservation	64	0.28	0.31	0.61	0.55	0.8	0.77
desertification	88	0.26	0.28	0.7	0.67	0.77	0.74
greenhouse effect	744	0.23	0.24	0.68	0.6	0.81	0.7
acid rain	237	0.3	0.31	0.73	0.62	0.76	0.68
rachel carson	24	0.29	0.29	0.75	0.67	0.75	0.67
soil erosion	1170	0.1	0.27	0.4	0.37	0.59	0.56
overfishing	71	0.18	0.18	0.51	0.48	0.56	0.55
X protection	350	0.19	1	0.43	0.38	0.51	0.45
dust bowl	116	0.07	0.17	0.27	0.2	0.47	0.38
gaia	106	0.26	0.26	0.45	0.29	0.54	0.36
protection of X	337	0.08	1	0.37	0.26	0.44	0.36