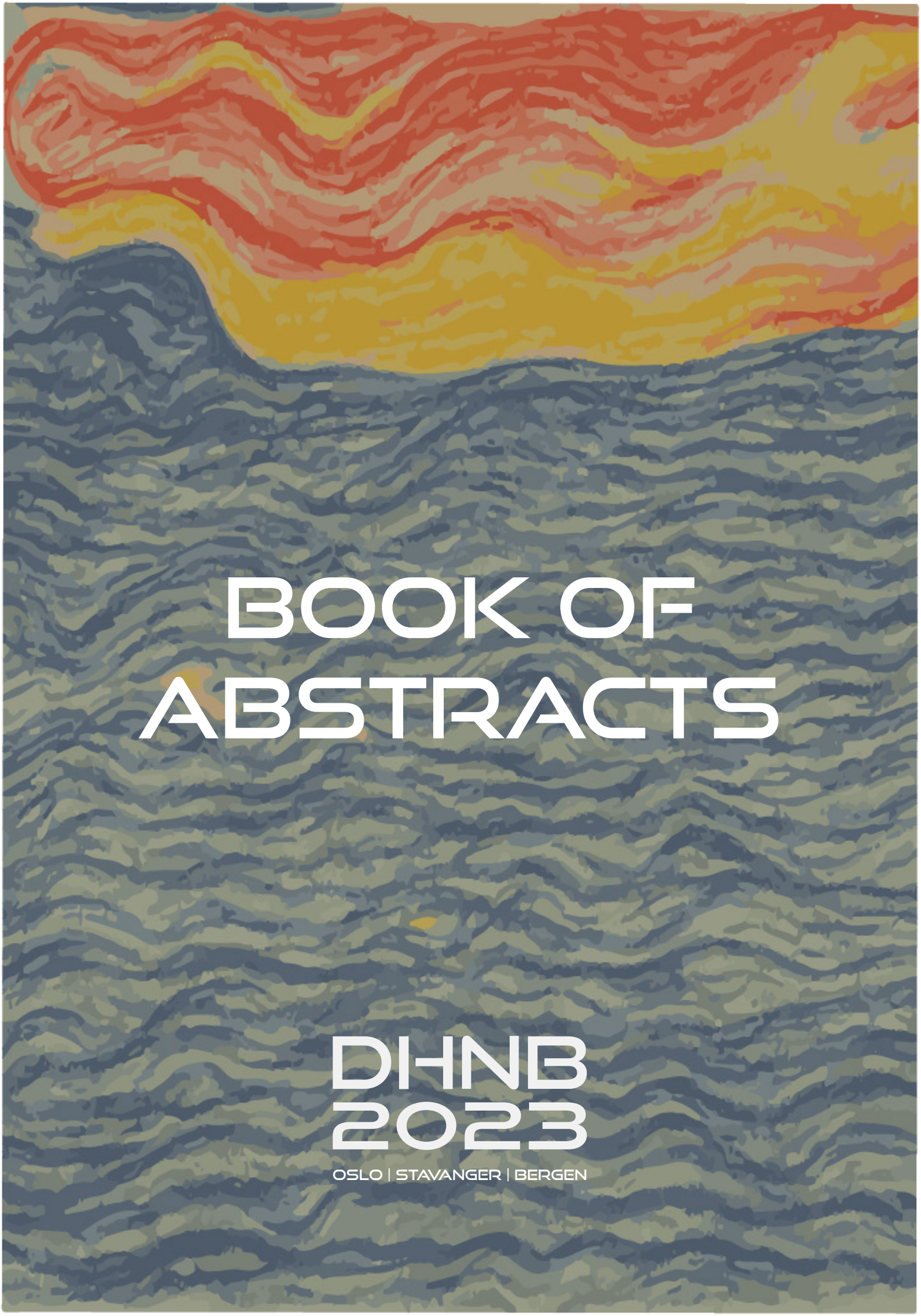

BOOK OF
ABSTRACTS
DHNB
2023
OSLO | STAVANGER | BERGEN

DHNB2023 Book of Abstracts

Book of Abstracts

Sustainability: Environment, Community, Data

Digital Humanities in the Nordic and Baltic Countries

Seventh Conference

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

Edited by Sofie Gilbert
and Annika Rockenberger

UNIVERSITETET I OSLO

ISBN 978-82-8037-200-0
DOI 10.5281/zenodo.7670464

Conference website https://dhnb.eu/conferences/dhnb2023/

Cover Design by Sofie Gilbert and Annika Rockenberger.

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

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

Contents

Acknowledgements xiii
Preface xv
Organizers xvii
Program Committee xix
Colour Legend xxi

Keynotes **1**
Building a Sustainable Research Infrastructure: The ELMCIP Electronic Literature Knowledge Base (*Scott Rettberg*) 3
Forecasting Sustainability: Speculative Ecologies at Work in Digital Humanities, Environmental Humanities, and Artificial Intelligence (*Lisa Swanstrom*) 4
Open Science for Enabling Reproducible, Ethical and Collaborative Research: Insights from the Turing Way (*Malvika Sharan*) 6

Long Papers **7**
Automated Coding of Danish Causes of Death 1861-1911. How Far Can We Get With String Similarity? (*Louise Ludvigsen, Mads L. Perner, Hilde L. Sommerseth, Bjørn-Richard Pedersen, Trygve Andersen, Lars Ailo Bongo, Rafael N. Cañadas, Anders Sildnes, Nikita Shvetsov*) . . 9
Benign Structures. The Worldview of Danish National Poet, Pastor, and Politician N.F.S. Grundtvig (1783-1872) (*Katrine Frøkjær Baunvig, Kristoffer Laigaard Nielbo*) 11
Building a Set of Digital Corpora and Data Sets to Examine the Discursive Representation of Nature and Environmental Change (*Gisle Andersen, Anje Müller Gjesdal, Knut Hofland, Marita Kristiansen*) 13
Building and Serving the Queerlit Thesaurus as Linked Open Data (*Arild Matsson, Olov Kriström*) 15
Chasing the Model. Experimenting With Training a Neural Network to Recognise Text in a Multi-Language and Multi-Authored Handwritten Document Collection (*Carlotta Capurro, Vera Provatorova, Sven Dupré, Marieke Hendriksen, Evangelos Kanoulas*) 18
CLARIN-DK: Supporting the Production and Distribution of FAIR Data – Achievements and Open Issues (*Costanza Navarretta, Dorte Haltrup Hansen*) 20
Community and Interoperability at the Core of Sustaining Image Archives (*Ulrike Felsing, Peter Fornaro, Max Frischknecht, Julien Antoine Raemy*) 22

A Complex Philosophical Œuvre and its Complex User Community: The Case of the Wittgenstein Archives Bergen (*Nivedita Gangopadhyay, Sebastian Sunday Grève, James Matthew Fielding, Alois Pichler*) . . 24
The Cultural Imaginary of Terrorism: Close and Distant Readings of Political Terror in Swedish News and Fiction During the Cold War (*Michael Azar, Daniel Brodén, Mats Fridlund, Michael McGuire*) . . 26
Detection and Clustering of Printers' Marks to Reveal the Publisher Networks of 18th Century Books (*Ruilin Wang, Yann Ryan, Lidia Pivovarova, Mikko Tolonen*) 28
Developing a Thesaurus-Based Semantic Tagger for Danish (*Ross Deans Kristensen-McLachlan, Nicole Dwenger, Sanni Nimb*) 30
The Diachrony of the New Political Terrorism: Tracing Neologisms and Frequencies of Terror-related Terms in Swedish Parliamentary Data 1971–2018 (*Daniel Brodén, Mats Fridlund, Leif-Jöran Olsson, Magnus P. Ängsal, Patrik Öhberg*) 32
The Digital Lab as an Arena for Teaching and Outreach Activities Connected to the Special Collections at the University of Bergen Library (*Emma Josefin Ölander Aadland*) 34
Digital Witnesses. Representation of the Bucza Genocide on the Social Media Platforms (*Bartosz Hamarowski, Maria Lompe*) 36
Distant Reading the Climate: Digital Analysis of Weather Information in 19th Century Press (*Krister Kruusmaa*) 38
Embed, Detect and Describe (EDDe): A Framework for Examining Events in Complex Sociocultural and Historical Data (*Melvin Wevers, Jan Kostkan, Kristoffer Laigaard Nielbo*) 40
Engineering Terrorismmindedness: A Scientometric Study of the 9/11-effect on STEM Research, 1989-2013 (*Mats Fridlund, Gustaf Nelhans*) . . . 42
Exploring Latvian Twitter Eaters' Food-Related Sentiment in Different Weather Conditions and in Relation to Meat (*Maija Kāle, Matīss Rikters*) 44
Exploring the Stability of Political Rhetoric in Finnish Parliamentary Debates Using Deep Learning (*Otto Tarkka, Kimmo Elo, Filip Ginter, Veronika Laippala*) 46
Facilitating Data Re-use by Better Understanding Paradata Outputs (*Olle Sköld, Jessica Kaiser, Lisa Andersson, Isto Huvila, Ying-Hsang Liu*) 48
Finding Historical Discourse on Natural Environment: Australian Newspapers 1900-1990 (*Peeter Tinits*) 51
The Future of Food Computing: Deepening the Scope by Broadening the Network (*Maija Kāle, Ramesh Jain*) 53
Generative Historicity: AI Image Synthesis as a Tool for Exploring Historicity and Historiography (*Per Gunnar Israelson, Matts Lindström*) 55
How Dark is Dark Souls? Applying Computer Vision to Analyze Video Game Walkthroughs (*Thomas Schmidt, Pascal Lindemann, Maximilian Huber*) 57

The Laborious Cleaning: Acquiring and Transforming 19th-Century Epistolary Metadata (*Senka Drobac, Johanna Enqvist, Petri Leskinen, Muhammad Faiz Wahjoe, Heikki Rantala, Mikko Koho, Ilona Pikkanen, Iida Jauhiainen, Jouni Tuominen, Hanna-Leena Paloposki, Matti La Mela, Eero Hyvönen*) 59

Managing Digital Humanities Data and Collections: The Records Continuum Model and the Collections of the Meertens Institute (*Douwe Arjen Zeldenrust*) 61

Nature and Culture in the Age of Environmental Crisis: Digital Analysis of a Global Debate in The UNESCO Courier, 1948-2011 (*Benjamin G. Martin, Fredrik Norén*) 63

Perspectives on Sustainable Dislocated Digital Research Resources (*Andrea Gasparini, Tom Gheldof*) 65

Picturing Swedish Women's History: Digitizing Photographs from the KvinnSam Archives (*Rachel Laura Pierce*) 67

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*) 69

Results From Rough Data? Using Multi-Dimensional Register Analysis to Study Scottish Enlightenment Historical Writing (*Aatu Liimatta, Yann Ryan, Tanja Säily, Mikko Tolonen*) 71

(R)Unicode: Encoding and Sustainability Issues in Runology (*Elisabeth Maria Magin, Marcus Smith*) 73

The SRDM Methodology for Sustainable Semantic Infrastructure: The Case of the Wittgenstein Archives Bergen and Its Related Resources (*James Matthew Fielding, Alois Pichler*) 75

Storage Over Rendition. Towards a Sustainable Infrastructure in the Digital Textual Heritage Sector (*Katrine Frøkjær Baunvig, Per Møldrup-Dalum, Krista Stinne Greve Rasmussen, Kirsten Vad*) 77

Stories of Sustainability (*Mareike Schumacher, Evelyn Gius, Itay Marienberg-Milikowsky*) 79

A Sustainable West? Analyzing Clusters of Public Opinion over Time and Space in a Collection of Multilingual Newspapers (1999-2018) (*Elena Fernandez Fernandez, Germans Savcisens*) 81

Tracing the Digital Plant Humanities: Narratives of Botanical Life and Human-Flora Relationships (*Paul Arthur, John Charles Ryan*) 83

Tracing the Proliferation of Socialist Realism Doctrine in Latvian Periodicals: Case Study of "Literature and Art" and "The Flag" (*Anda Baklāne, Valdis Saulespurēns*) 85

Transliteration Model for Egyptian Words (*Heidi Jauhiainen, Tommi Jauhiainen*) 87

Understanding Without Knowing: Livonian Intangible Cultural Heritage (*Valts Ernštreits*) 89

Untapped Data Resources: Applying NER for Historical Archival Records of State Authorities (*Venla Räsänen, Tanja Välisalo, Ida Toivanen, Jari Lindroos, Antero Holmila, Jari Ojala*) 91
Vectors of Violence: Changing Conceptualizations of Terror, Terrorism and Våldsbejakande Extremism in Swedish Parliamentary Data, 1971–2018 (*Magnus P. Ängsal, Daniel Brodén, Mats Fridlund, Leif-Jöran Olsson, Patrik Öhberg*) 93
Våran Klubb, Barnvaktsklubben or The Baby-Sitters Club: The Data-Sitters Club as an International Community (*Agnieszka Backman, Quinn Dombrowski*) 95
Wikidata for Authority Control: Sharing Museum Knowledge With the World (*Alicia Fagerving*) 97

Show-and-Tells **99**
Air Pollution Affecting the Speech Complexity of Finnish Members of Parliament (*Anna Kristiina Ristilä*) 101
Collecting and Sharing References in Christian-Muslim Religious Encounters With the OTRA Framework: An Assessment and a Roadmap (*Jacob Langeloh*) 103
Creating a Crowdsourcing System and Community to Collect Information on Local Everyday Objects (*Sakiko Kawabe*) 104
Digitisation of Printed Books in Old Latvian Orthography for the Preservation and Sustainability of Cultural Heritage: Workflow and Methodology (*Karina Šķirmante, Silga Sviķe*) 105
Dos and Don'ts of Building a Pan-European Biographical Knowledge Graph: Statistical Analysis of the InTaVia-Platform (*Matthias Schlögl, Joonas Kesäniemi, Jouni Tuominen, Victor de Boer, Go Sugimoto, Carla Ebel*) 106
Environmental Concerns in COVID-19 Vaccine Discussions on Twitter: Between Science Enthusiasm and Science Denial (*Jana Sverdljuk, Bastiaan Bruinsma*) 107
Making Modern Drama: Change Detection in the Information Space of Henrik Ibsen's Writing (*Ellen Rees, Frida Hæstrup, Kristoffer Laigaard Nielbo*) 108
An OCR Pipeline for Transforming Parliamentary Debates into Linked Data: Case ParliamentSampo – Parliament of Finland on the Semantic Web (*Senka Drobac, Laura Sinikallio, Eero Hyvönen*) 110
Phytobibliography: Building a Digital Database of Plants in Scandinavian Picturebooks for Children (*Beatrice G. Reed*) 112
Reconstructing MultiTorg: Archaeological Approaches to Digital Artefacts of the Historical Web (*Jon Carlstedt Tønnessen*) 113
Search and Exploring Correspondence: correspSearch (*Stefan Dumont, Sascha Grabsch, Ruth Sander*) 115
Some Nobel Laureates Are More Coherent Than Others: Measuring Literary Quality in a Corpus of High Prestige Contemporary Literature (*Yuri Bizzoni, Pascale Feldkamp Moreira, Ida Marie Lassen, Kristoffer Laigaard Nielbo, Mads Rosendahl Thomsen*) 116

Towards Reusable Aggregated Biographical Research Data: Provenance and Versioning in the InTaVia Knowledge Graph (*Joonas Kesäniemi, Matthias Schlögl, Jouni Tuominen, Victor de Boer, Go Sugimoto*) . . 117
Transforming Linguistically Annotated Finnish Parliamentary Debates Into the Parla-CLARIN Format (*Minna Tamper, Laura Sinikallio, Jouni Tuominen, Eero Hyvönen*) 118
Virtual Lab at the National Library of Estonia (*Peeter Tinits, Urmas Sinisalu*) 119
Visualising the Cuneiform Corpus: Results of the Project Geomapping Landscapes of Writing (GLoW) (*Seraina Nett, Nils Melin-Kronsell, Carolin Johansson, Gustav Ryberg Smidt, Rune Rattenborg*) 120
The Words of Climate Change: TF-IDF-Based Word Clouds Derived From Climate Change Reports (*Maria Skeppstedt, Magnus Ahltorp*) 121

Workshops **123**
Challenges of Long-Term Sustainability of DH Projects: The LAM (Libraries, Archives and Museums) Perspective (*Olga Hołownia, Helena Byrne, Grace Bicho*) 125
Cross-University Collaboration in Digital Humanities and Social Science (DHSS) & Digital Humanities & Cultural Heritage (DHCH) Education (*Jonas Ingvarsson, Ahmad Kamal, Koraljka Golub, Isto Huvila, Olle Sköld, Anna Foka, Marianne Ping Huang, Mikko Tolonen*) 127
Exploring Digital Tools and Platforms for Individual Research of History and Antiquity (*Victoria G. D. Landau, Sarah Siegenthaler*) 130
KUB Datalab's Digital Humanities Workshop (*Lars Kjær*) 132
The Norwegian Web Archive: Searching and Examining the Web of the Past (*Jon Carlstedt Tønnessen*) 133
To Serve Them All – Web Accessibility in Digital Humanities (*Tone Merete Bruvik*) 134

Author Index **139**

Country Index **143**

Institution Index **145**

Acknowledgements

The DHNB2023 organizers would like to express their gratitude for the substantial financial and continuous moral support from Cecilia Ekström, Head of the Humanities and Social Sciences Library at the University of Oslo Library. We thank Karin Cecilia Rydving, Section Head of Education and Research Support at the University of Bergen Library and Dolly and Finn Arne Jørgensen, Co-directors of The Greenhouse Center for Environmental Humanities at the University of Stavanger, for the support they have given during the planning and organization of the conference and the hosting of the on-site keynotes in Bergen and Stavanger. We thank the DHNB Board for its guidance and the provision of the DHNB Conference Handbook, even in its draft form. We thank the Program Committee for reviewing all submitted papers and presentations and supervising the assembly of the conference program. Lastly, we thank Gisela Attinger, Matthew Good, Astrid Anderson, Federico Aurora, Anne Sæbø, Oda Amalie Rosenkilde, Elisa Pierfederici, Olga Hołownia, Naomi Yabe Magnussen, Jenny Ostrop, Randi Cathinka Neverdal, Runhild Seim, Mads Baklien, Ragnhild Sundsbak, Pål Magnus Lykkja, Serena Norlemann Baldari, Agata Bochynska, Hilde Kristin Hem, Anne-Gry Skonnord, Johanne Christensen, Jenny Søbyskogen, Maylinn Hovengen Byrknes and Maria Lien for their help, assistance, and encouragement.

The DHNB2023 Organizers, Annika Rockenberger, Sofie Gilbert, Juliane Marie-Thérèse Tiemann, Finn Arne Jørgensen.

Preface

The 7th annual Digital Humanities in the Nordic and Baltic Countries Conference (DHNB2023) is held online from 8–10 March 2023. The DHNB conferences focus on research, education and communication in the interdisciplinary digital humanities and social sciences field in the Nordic and Baltic regions and beyond. DHNB2023 explores the many facets of Sustainability in the Digital Humanities and Social Sciences with a particular focus on Environment, Community, and Data.

Environment
The Digital Humanities do not stand at a distance from the environmental challenges facing the planet. In 2014, Bethany Nowviske challenged DH scholars and practitioners to consider the place of the field in the Anthropocene. What responsibilities do we have as the world around us burns, dries, drowns, and changes before our eyes as species go extinct and ways of life end? How do DH projects and practices depend on unsustainable systems and mindsets? How do the unequal consequences of environmental challenges influence what research gets done in DH, and who can contribute? How can the field 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 organisations 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 centres 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? Moreover, how can we shape and create its future success together? Sustainable communities are places where people of diverse backgrounds and perspectives feel welcome and safe, where every group and every member has a say in the decision-making process, and where intellectual prosperity is shared. What does this mean for DHNB now and in the future?

Data
The primary source material for humanists has many data formats, and research is becoming increasingly digital and, in many cases, only exists in digital form. As increasing digitisation 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 in terms of 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?

DHNB2023 in Numbers

The Call for Submissions received 85 proposals for the formats Long Paper, Show-and-Tell Presentation, Panel, and Workshop. The Program Committee selected 68 submissions for the conference after a single-blind peer-review process: 44 Long Papers, 20 Show-and-Tell Presentations, and six workshops. The submissions span a wide array of topics within Digital Humanities, from 3D modelling to web research. For the program, we have gathered the presentations into three thematic tracks – Environment, Data, and Community – and two open tracks, focusing on either specific DH methods or research objects. The keynotes feature each day's central theme and conclude the day.

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 as shown in the Country Index.

We thank all presenters and participants for making the conference a rich experience and a meaningful way of engaging with the Nordic and Baltic as well as the international DH community.

The Organizers, Oslo/Bergen/Stavanger, 22nd February 2023

Organizers

Annika Rockenberger, University of Oslo Library, Norway

Sofie Gilbert, University of Oslo Library, Norway

Juliane Marie-Thérèse Tiemann, University of Bergen Library, Norway

Finn Arne Jørgensen, University of Stavanger, The Greenhouse Center for Environmental Humanities, Norway

UNIVERSITY
OF OSLO

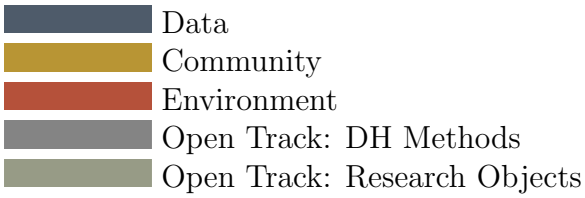
UNIVERSITY
OF BERGEN

Program Committee

Annika Rockenberger (Chair), University of Oslo, Norway

Emma Aadland, University of Bergen, Norway
Anda Baklāne, National Library of Latvia, Latvia
Peder Gammeltoft, Norwegian Language Collections, Norway
Sofie Gilbert, University of Oslo, Norway
Finn Arne Jørgensen, University of Stavanger, Norway
Lars Kjær, The Royal Danish Library, Denmark
Costanza Navarretta, University of Copenhagen, Denmark
Torsten Roeder, University of Würzburg, Germany
Juliane Marie-Thérèse Tiemann, University of Bergen, Norway
Jurgita Vaičenonienė, Vytautas Magnus University, Lithuania

Colour Legend



Keynotes

Building a Sustainable Research Infrastructure: The ELMCIP Electronic Literature Knowledge Base

Scott Rettberg

University of Bergen, Norway

9 Mar 2023
19:00-20:00

The ELMCIP Electronic Literature Knowledge Base is an open-access research database for the field of electronic literature, initially developed in 2010 for the HERA-Funded collaborative research project Electronic Literature as a Model of Creativity and Innovation in Practice. The database has now been in continuous production for 13 years. The basic premise of the Electronic Literature Knowledge Base model is that it considers a field as a network composed of human and non-human actors, objects, and events. The literary artifact is inseparable from the network in which it is produced, disseminated, and post-processed. The database documents individual objects, but even more importantly, is based on a knowledge model that accounts for and makes available for study the relations between them. This talk will describe the process and challenges of developing and maintaining the project up until now, and the changes we plan to make to the database as we adapt it for new uses in the Center for Digital Narrative, a Norwegian Center of Research Excellence that is launching in August 2023.

Forecasting Sustainability: Speculative Ecologies at Work in Digital Humanities, Environmental Humanities, and Artificial Intelligence

Lisa Swanstrom
University of Utah, United States of America

8 Mar 2023
19:00-20:00

In the sixth book of the Old Testament, the prophet Joshua prays for a miracle to ensure that the Israelites will prevail over their adversaries. God grants his request, and a total eclipse of the sun ensues: "And the sun stood still, and the moon stopped, until the nation took vengeance on their enemies" (Joshua 10:13). In sixth chapter of Mark Twain's A Connecticut Yankee in King Arthur's Court, a similar "miracle" occurs, which saves the hide of the novel's narrator, Hank, who faces execution unless he can deliver on his promise of an eclipse: "as sure as guns, there was my eclipse beginning! ... The rim of black spread slowly into the sun's disk, my heart beat higher and higher..." (41). Although separated by large swaths of time and at farcical odds in terms of both audience and intent, these two moments from literary history help illustrate an important lesson about Artificial Intelligence.

In both works, the performance of divination is captivating for its violation of natural law. That is, the sun's disappearance defies causality, cementing Joshua's connection to God in the Bible and saving Hank's bacon in Twain. If we take the texts at face value, however, only Joshua's eclipse is "miraculous." He calls upon God; God answers. In contrast, Twain's novel depends upon narrative, rather than divine, machinations. Hank has somehow been transported back in time and is scheduled to be burned at the stake. Because he is modern man of the late nineteenth century, however, and an engineer to boot, he knows a thing or two about the laws of physics and is thus able to predict the eclipse and stay his execution. Both events, however, are staged as instances of prophecy—and opportunity—from which both men profit.

Similarly, discussions of AI's ability to parse human communication, through Natural Language Processing (NLP) and its ilk—Machine Learning, Neural Networking, Sentiment Analysis, Data Mining, etc.—often focus on what appears to be its miraculous capacity for prognostication, if not the inevitability of a computational Singularity.[1] In point of fact, however, it is Twain's narrator, shifty and self-motivated as he is, who provides a more accurate model for understanding such technology. In the case of Joshua, the prediction leaps toward the future, un- tethered by rational evidence. In the case of his less illustrious counterpart, the success depends equally upon the retrospective assessment of the past and the occlusion of this knowledge from the present. Data forecasting is not prophecy. Rather, it is a science of extrapolation that depends upon probability and statistical analysis. This might

[1] Advertising, in particular, draws from cliched images of AI from a science fictional past. Consider the following animated gif, which combines the clunky body of Robby the Robot with a glowing globe that looks quite a bit like the Wikipedia logo. This image accompanies an article in the Intercept about AI, surveillance, and social media (Biddle). A series of images from KnowledgeNile, a company specializing in "technology based content" and machine learning strategies for marketing, provides an apt complement. Here KnowledgeNile takes clear (shameless?) inspiration from (slightly) more contemporary works of sf, including, from left to right, Stephen Spielberg's AI, Robert Longo's Johnny Mnemonic (or possibly Tron), and the Wachowskis' Matrix (and/or Björk's "All Is Love").

seem like a banal assessment, but it warrants consideration. Statistics do not merely anticipate outcomes. They also have the capacity to shape what they purport to measure, control what is to be counted in their reckoning, and elide what is not. Statistics form an effective speculative epistemology, to be sure, but also a sneaky and redacted one.

Considering the increasing demand for NLP and other means of text-based forensic technology, it is worth our time to examine their appearance in the cultural imagination against their technological reality.[2] This paper proceeds by doing just this. In the first section, I identify a common yet misleading story that circulates around AI in order to outline an alternative aesthetic genealogy. Secondly, and in dialog with Amitav Ghosh's assertion that "Probability and the modern novel are in fact twins... born at about the same time, among the same people, under a shared star..." (16), I discuss the importance of re-framing AI within literary studies and the Digital Humanities in a way that confronts its statistical underpinnings. I conclude by offering a different, more playful approach to Natural Language Processing within literary scholarship that is committed to an expansive, material, and environmentally responsible concept of intelligence—artificial or otherwise.

References

Biddle, Sam. "Facebook Uses Artificial Intelligence to Predict Your Future Actions for Advertisers, Says Confidential Document." The Intercept 13 April 2018.

Columbus, Louis. "Roundup of Machine Learning Forecasts and Market Estimates, 2020." Forbes 19 Jan 2020.

Ghosh, Amitav. The Great Derangement: Climate Change and the Unthinkable. New York: Penguin, 2016.

Joshua 10:13. New International Version. BibleHub.

Twain, Mark. A Connecticut Yankee in King Arthur's Court. 1890. Oxford UP, 1998.

[2] A series of images from KnowledgeNile, a company specializing in "technology based content" and machine learning strategies for marketing, provides an apt complement. Here KnowledgeNile takes clear (shameless?) inspiration from (slightly) more contemporary works of sf, including, from left to right, Stephen Spielberg's AI, Robert Longo's Johnny Mnemonic (or possibly Tron), and the Wachowskis' Matrix (and/or Björk's "All Is Love").

Open Science for Enabling Reproducible, Ethical and Collaborative Research: Insights from the Turing Way

Malvika Sharan

The Alan Turing Institute London, Great Britain

10 Mar 2023
19:00-20:00

As researchers, we make complex choices around project design and decisions throughout the lifecycle of our research. We are expected to ensure that our research objects are easily accessed, openly examined and built upon by others in future work. Although open and transparent reporting helps to make sure that scientific work can be trusted, we must also integrate considerations of the societal and ethical implications of our work. Especially when these considerations impact people's lives. Furthermore, reproducibility alongside open research practices is important for enabling independent verification of research methods, underlying data, analysis code and workflows. All these require an understanding of research best practices and skills that are often not widely taught or explored among academic researchers.

In this talk, I will discuss open science as a framework to ensure that all our research components can be easily accessed, openly examined and built upon by others. I will introduce The Turing Way - an open source, open collaboration and community-driven guide to reproducible, ethical and inclusive data science and research. Drawing insights from the project, I will share best practices that researchers should integrate to ensure the highest reproducible and ethical standards from the start of their projects so that their research work is easy to reuse and reproduce at all stages of development. All attendees will leave the talk understanding the many dimensions of openness and how they can participate in an inclusive, kind and inspiring open source ecosystem as they collaboratively seek to improve research culture. All questions and contributions are welcome at the GitHub repository: https://github.com/alan-turing-institute/the-turing-way.

Long Papers

Automated Coding of Danish Causes of Death 1861-1911. How Far Can We Get With String Similarity?

Louise Ludvigsen[1], Mads L. Perner[1, 2], Hilde L. Sommerseth[3], Bjørn-Richard Pedersen[3], Trygve Andersen[3], Lars Ailo Bongo[3], Rafael N. Cañadas[3], Anders Sildnes[3], Nikita Shvetsov[3]

[1]University of Copenhagen, Denmark; [2]Danish National Archives, Denmark; [3]UiT The Arctic University of Norway, Norway

10 Mar 2023
12:30 - 14:00

Historical cause of death data have been central to some of the most influential studies of the modern mortality decline in the 19th and 20th century, including Omran's "The Epidemiologic Transition" and McKeowns "The Modern Rise of Populations". However, their use of published aggregated statistics as their main source, pre-classified by contemporaries, has been heavily criticized, and it is questioned what can really be gained from this type of data. Following years of work by archives and volunteers on digitizing historical sources, researchers now have access to unprecedented datasets and materials in both size and detail. In historical mortality studies, the digitization of individual-level causes of death, from sources such as parish registers, vaccination protocols, death certificates and burial records, have been game-changing. With individual-level causes of death, we have the opportunity for a much more in-depth analysis of the factors that influence mortality in a given population.

One such dataset is the Copenhagen Burial Register, which has been digitized and transcribed for the period 1861-1911 by the Copenhagen City Archives. It contains over 300,000 individual burials and more than 10,000 unique strings of causes of death. This means that for the first time, it is now 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? Manual coding would be the go-to method for historians, but as this is a very time-consuming process and one that requires considerable domain expertise, it is a method that does not scale well and is not sustainable in the long run. One possible solution is to supplement manual coding with coding done by automated string similarity matching. However, the quality of this approach has not been tested.

This paper aims to see how far we can get with automated coding based on string similarity. We do this by applying a set of simple but automated string similarity algorithms that code our cause of death data to DK1875, a contemporary coding and classification system from 19th-century Denmark. 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 that of a manual (historian) coder with extensive domain expertise.

Our preliminary results show that a minimum-effort algorithm coded approximately half of the causes of death correctly compared to the manually coded dataset. However, with some efforts to standardize the data and accommodate differences in spelling, an accuracy of 60-70% is within reach. We discuss what consequences this level of accuracy would have for analysis, by comparing with the manually coded dataset.

References

Omran. (2005). The Epidemiologic Transition: A Theory of the Epidemiology of Population Change. The Milbank Quarterly, 83(4), 731–757.

McKeown, T. (1976). The modern rise of population. Edward Arnold.

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

Katrine Frøkjær Baunvig, Kristoffer Laigaard Nielbo
Aarhus University, Denmark

10 Mar 2023
14:30-16:00

In Denmark N.F.S. Grundtvig (1783-1872) plays the dual role of Church Father and Founding Father. Through his highly popular hymns on the one hand, and his secular song lyrics on the other, he has provided a world-affirming poetic universe for Danish mainstream religious life and Danish cultural mentality respectively. That his worldview is fundamentally positive becomes evident when plotting the semantic tissue conjoining 'Earth', 'Heaven', and Hell'. These three terms represent the most significant ontic domains in Grundtvig's thinking specifically, and in Christian history generally. The center of Grundtvig's universe is the heavenly 'Sun' radiating benign, divine energy to the human realm, the 'Earth'. That is: there is a strong and sustainable integration of the godly and the human sphere. 'Hell' enjoys no such connection. In fact, 'Hell', and the semantic neighbour 'Death', are strikingly isolated in Grundtvig's worldview.

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 the centuries leading up to the Common Era's beginning (Bellah 2017). It is a classical argument that this development undergirded anthropocentrism and human exploitation of natural resources (e.g., White 1967). 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 centuries. 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' is a more plausible semantic framework than 'damnation'; a poetic platform for future sustainable developments in Denmark?

Design and Data

Embeddings of 'Heaven', 'Earth' and 'Hell' in Grundtvig's Works
This study combines simple neural embeddings and graph theory to represent the worldview arising from Grundtvig's 1068 publications in their tokenized, lemmatized, 'algorithmified' avatar. We demonstrate 1) that the center of Grundtvig's semantic worldview is the lifegiving, heavenly, divine sun; 2) that Hell is an insulated, arid domain.

This study is based on the digital scholarly edition Grundtvig's Works enriched by philologists' strenuous mark-up and thus offering a clean, reliable, and flexible corpus-material open to comprehensive and hermeneutically complex explorations (Rasmussen et al. 2022). In order to plot Grundtvig's semantic worldview, we have made the Grundtvig-algorithm compute the distance between a set of so-called seed terms (Heaven [Himmel]; Earth [Jord]; Hell [Helvede]) and the corpus lexicon. Further: for each seed, the algorithm has extracted terms with the shortest distance (primary associations) as well as terms with the shortest distance to the

primary associations (secondary associations). The distance between all terms (i.e., seeds, primary and secondary associations) has been computed and terms have been connected based on their distance under a given threshold. Finally, semantic clusters have been extracted using the Louvain method (Blondel et al. 2008) and the visualization is represented generated with terms as nodes and thresholded distances as edges.

References

Bellah, Robert. 2017. Religion in Human Evolution. From the Paleolithic to the Axial Age, Harvard University Press.

Blondel, Guillaume, Lambiotte & Lefebvre. 2008. "Fast unfolding of communities in large networks", Journal of Statistical Mechanics: Theory and Experiment, 1-12.

Rasmussen, K.S.G., K.S. Ravn, J. Tafdrup & K.F. Baunvig. 2022. "The Case for Scholarly Editions", DHNB 2022 Proceedings.

White Jr., L. 1967. "The historical roots of our ecological crisis", Science, 155(3767), 1203-1

Building a Set of Digital Corpora and Data Sets to Examine the Discursive Representation of Nature and Environmental Change

Gisle Andersen[1], Anje Müller Gjesdal[2], Knut Hofland[1], Marita Kristiansen[3]
[1]NORCE Norwegian Research Centre, Norway; [2]Østfold University College, Norway;
[3]University of Bergen, Norway

8 Mar 2023
14:30-16:00

As environmental change accelerates and increasingly impacts more aspects of our lives, understanding how people make sense of these phenomena becomes ever more important. How is knowledge about nature represented and negotiated in order to allow for public debate and political action?

This paper presents results from Changing Nature: A lexical and argumentative analysis of public debates on nature, a research project funded by the Research Council of Norway that examines the relationship between nature and language change in Norway in the time period 1998-2019. The analytical focus of the project is on the relationship between language change and conceptual change in the environmental domain and specifically of the contribution of neology to public debate (for an overview, see Andersen & Gjesdal, 2020; Gjesdal & Kristiansen, 2021). In order to investigate this topic, the project has developed a set of relevant digital corpora and data sets.

This paper describes the collection and construction of these resources based on methodologies developed previously by members of the research team (Andersen & Hofland, 2012; Losnegaard et al., 2013). The first data set is the Naturen corpus. Naturen is a popular science journal which was founded in 1877, and it is the oldest popular science journal in Norway. The second data source is a data set of NOU reports which has been compiled in the context of the Changing Nature project. The data set consists of all the NOUs from 1998-2017, 669 reports in total. The NOUs cover a range of issues that are deemed to be timely and topical by decision makers, as they are written by expert committees at the request of a ministry. Thus, they also discuss crucial issues related to nature and the environment in this period. Finally, we present a word list of environmental neologisms derived from the data sets.

The paper will present the resources that have been developed in the project as well as a pilot study to demonstrate the interest in combining different data sets to examine the interaction of lexical change and language change in the environmental domain. We compare neologisms in the two corpora from the domains of climate change and biodiversity with a view of how they differ quantitatively and qualitatively (collocations, patterns of neology formation).

References

Andersen G and Gjesdal AM (2020) Karbonsnakk – hva snakker vi om når vi bruker begrepet "karbon"? Nytt norsk tidsskrift 37(2): 163-178.

Andersen, G., & Hofland, K. (2012). Building a large corpus based on newspapers from the web. In G. Andersen (Ed.), Exploring Newspaper Language - Using the web to create and investigate a large corpus of modern Norwegian (pp. 1-30). John Benjamins Publishing.

Gjesdal AM and Kristiansen M (2021) Communicating Natural Events. Emerging Terminology across Corpora In: V. Delavigne V and De Vecchi DD (eds) Termes en discours. Entreprises et organisations. Paris: Presses Sorbonne Nouvelle.

Losnegaard, G. S. et al. (2013). Linking Northern European infrastructures for improving the accessibility and documentation of complex resources. Proceedings of the workshop on Nordic language research infrastructure at NODALIDA 2013. NEALT Proceedings Series 20 / Linköping Electronic Conference Proceedings 89: 44–59.

Building and Serving the Queerlit Thesaurus as Linked Open Data

Arild Matsson, Olov Kriström
University of Gothenburg, Sweden

This paper will describe and discuss a digital infrastructure supporting the development and implementation of a specialized thesaurus for subject indexing of literature using Linked Open Data (LOD).

9 Mar 2023
12:30-14:00

Background

The Queerlit project aims to identify and subject-index queer, Swedish fiction. Literary scholars deem existing applicable thesauri unsuitable for this purpose, as these fail to express the topics and nuances required to appropriately navigate the bibliography (Hansson 1999, Campbell 2001 and 2004, Olson 2002, Samuelsson 2008, Bates & Rowley 2011, Christensen 2011, Adler 2017). Thus, as a part of the Queerlit project, librarians and literary scholars are together constructing a new thesaurus: the Queer Literature Indexing Thesaurus (QLIT).

QLIT is based on the Homosaurus, a large thesaurus with similar goals as the present one (Homosaurus n.d.). Homosaurus was translated, adapted to a Swedish context and then adjusted in iterations, in conjunction with discussion and ongoing indexing work.

As a participant in the project, the National Library of Sweden (Kungliga biblioteket, KB) ensures that the indexing work and the thesaurus are recorded in their bibliographical database, Libris. Libris uses the BIBFRAME data framework (McCallum 2017, KB n.d.) which uses linked data with the Resource Description Framework (RDF).

Method

Construction of the thesaurus was realized with rather minimalistic technical solutions. The work can be split into three parts:

Adaptation of the Homosaurus

- Downloading the Homosaurus in Turtle RDF format
- Translating, revising, removing and adding terms and relations

Data normalization

- Validating and parsing the Turtle definition of each term
- Merging all term data and reporting consistency errors
- Automatic enhancement of the data
- Committing the resulting RDF graph to version control

This workflow is almost fully scripted in Python.

Web server

Data is served in two variants with a light-weight web server application:

- Re-serializing the full data back into Turtle format for importing into Libris
- Several endpoints serving simple JSON structures for the separately developed Queerlit website, which features a user-friendly tool for exploring the thesaurus

Result

With this infrastructure in place, continual work on the thesaurus is a semi-automatic process. The author edits Turtle files in a text editor. To publish changes, a sequence of commands are issued to execute the automated parts of the process: running the normalization script, checking the diff, committing to version control, pushing to a production server and restarting the server application.

Discussion

Existing tools for working with RDF data and thesauri include editor interfaces, validators, visualizers and web servers. The decision not to use any of these was motivated partly by preference for plain-text editing and partly due to requirements being quite simple:

- The project member responsible for authoring data quickly got acquainted with the Turtle syntax used in the Homosaurus export, and found manual editing easier than evaluating and learning graphical RDF editors
- The import pipeline into Libris simply required a single RDF dump to be available on the web
- For visualization, there was already a public website in development, which would contain a technically and visually integrated thesaurus explorer

For a sustainable digital infrastructure, it is generally wiser to use or build upon existing and well-established alternatives than to build new software. The software built in this project was therefore kept to a minimum, and designed with respect to possible future adaptation to similar use-cases.

Authoring is performed directly in RDF syntax, which places a certain threshold on the digital competence of whoever is tasked with this responsibility. This could be an issue for long-term maintenance, but it is mitigated by the fact that the subsequent publication workflow still requires a developer to step in.

References

Bates, J. & Rowley, J. 2011."Social reproduction and exclusion in subject indexing: A comparison of public library OPACs and LibraryThing folksonomy." Journal of Documentation, vol. 67, no 3. 421-448.

Campbell, D. G. 2001. "Queer theory and the creation of contextual subject access tools for gay and lesbian communities." Knowledge Organization, vol. 27, no 3, pp. 122–31.

Campbell, D. G. 2004. “A Queer Eye for the Faceted Guy: How a Universal Classification Principle can be Applied to a Distinct Subculture.” Knowledge Organization and the Global Information Society: Proceedings of the 8th International Conference of the International Society for Knowledge Organization. London, U.K., 13-16 July 2004.

Christensen, B. 2011. “Interfiling intersex: How Dewey classifies intersex in theory and practice.” In Ellen Greenblatt (ed.) Serving LGBTIQ library and archives users: essays on outreach, service, collections and access, pp. 201–211. Jefferson, North Carolina: McFarland & Co.

Hansson, J. 1999. “Klassifikation, bibliotek och samhälle: en kritisk hermeneutisk studie av Klassifikationssystem för svenska bibliotek.” Diss. University of Gothenburg.

Homosaurus. n.d. “About.” Retrieved from homosaurus on 13 October 2022.

KB, Kungliga biblioteket. n.d. “Vad är Libris och XL?” Retrieved from libris/about on 14 October 2022.

McCallum, S. H. 2017. “BIBFRAME Development.” JLIS.it, vol. 8, no 3, pp. 71–85.

Olson, H. A. 2002. “The power to name: locating the subject representation in libraries.” Dordrecht; Boston, Mass.: Kluwer.

Samuelsson, J. 2008. “På väg från ingenstans: kritik och emancipation av kunskapsorganisation för feministisk forskning.” Diss. Umeå University.

Chasing the Model. Experimenting With Training a Neural Network to Recognise Text in a Multi-Language and Multi-Authored Handwritten Document Collection

Carlotta Capurro[1], Vera Provatorova[2], Sven Dupré[1], Marieke Hendriksen[3], Evangelos Kanoulas[2]

[1]Utrecht University, The Netherlands; [2]University of Amsterdam, The Netherlands; [3]The Royal Netherlands Academy of Arts and Sciences, The Netherlands

10 Mar 2023
12:30-14:00

This work aims at developing an optimal strategy to automatically transcribe a large quantity of uncategorised digitised archival documents when resources include handwritten text by multiple authors and in several languages. We present a comparative study to establish the efficiency of a single multilingual handwritten text recognition (HTR) model trained on multiple handwriting styles as opposed to using a separate model for every language. This approach allows us to automate the transcription of the archive, reducing manual annotation efforts and facilitating information retrieval. A good multilanguage model has several critical advantages. First, it allows documenting resources using entities already in the text,[1] reducing the interpretative efforts, thus granting more equitable access to resources. Second, it could be re-used on similar collections, reducing the investment of resources in training new AI models.[2] To train the model, we use Transkribus, a platform enabling AI-powered HTR.[3] We use the material from the personal archive of the Dutch glass artist Sybren Valkema (1916-1996).

Valkema was among the founders of the European branch of the Studio Glass movement that spread internationally from the US in the early 1960s, advocating the use of glass as an artistic medium.[4] The movement brought a revolution in glassblowing thanks to the invention of a studio furnace that allowed artists to become independent from industrial facilities.[5] To stimulate the free circulation of knowledge about glass-making techniques, Studio Glass artists established several courses on glassblowing in universities and art academies worldwide. During his life, Valkema combined glass blowing with an intense career as a teacher. He built the first studio glass furnace in Europe at the Rietveld Academy and developed a curriculum on glassmaking.

The Valkema archive is a valuable resource for studying the Studio Glass movement. It has been digitised, and research is being carried out to make it publicly accessible and easily searchable.[6] The archive contains over 103.000 pages documenting the artist's career, including teaching materials, letters, designs, descriptions of processes and many recipes for glass. All handwritten documents in the archive are scanned and stored as images. To make the data searchable, these images need to be converted into text using HTR. Current HTR methods require extensive manual annotation efforts to create training data: every language and every handwriting have unique features and ideally require the use of a separate model. Moreover, due to the lack of metadata about the language and author, we cannot automatically sort out the material for processing it with a specific model. To overcome this challenge, we create a multilingual, multi-hands model trained on samples from the Valkema archive and apply automatic post-correction to its output.

This contribution describes the methodology we developed to assess the quality of our multilingual and multi-hands model compared to using a separate model for

each language. Our experiments consist of four steps. First, for each of the four most common languages in the archive (Dutch, English, German and French), we selected and annotated a set of 50 documents. Second, every document is automatically transcribed using two different models:

- a monolingual model specific to the language of the document (expected to offer high-quality results);
- our multilingual model (expected to provide lower-quality results in all languages).

Third, the transcriptions produced using the multilingual model are processed with a post-correction algorithm to fix the transcription mistakes. Last, the accuracy of these results will be compared with the automatic transcription obtained using the corresponding monolingual models.

When the results of this experiment are satisfactory, the datafication of the archival collection will be speeded up, and the amount of human labour involved in the process will be reduced.

References

1 Vera Provatorova et al., 'Named Entity Recognition and Linking on Historical Newspapers: UvA.ILPS & REL At', 2020.

2 Aimee van Wynsberghe, 'Sustainable AI: AI for Sustainability and the Sustainability of AI', AI and Ethics 1, no. 3 (1 August 2021): 213–18, https://doi.org/10.1007/s43681-021-00043-6.

3 Sebastian Colutto et al., 'Transkribus. A Platform for Automated Text Recognition and Searching of Historical Documents', in 2019 15th International Conference on EScience (EScience) (2019 15th International Conference on eScience (eScience), San Diego, CA, USA: IEEE, 2019), 463–66, https://doi.org/10.1109/eScience.2019.00060.

4 Sybren Valkema and Kl Laansma, Sybren Valkema (Baarn: De Prom, 1994).

5 Joan Falconer Byrd and Harvey K. Littleton, Harvey K. Littleton - a Life in Glass: Founder of America's Studio Glass Movement (New York: Skira Rizzoli, 2011).

6 'Archief Sybren Valkema', RKD Nederlands Instituut voor Kunstgeschiedenis, accessed 2 January 2023, https://rkd.nl/nl/projecten-en-publicaties/projecten/265-archief-sybren-valkema.

CLARIN-DK: Supporting the Production and Distribution of FAIR Data – Achievements and Open Issues

Costanza Navarretta, Dorte Haltrup Hansen
University of Copenhagen, Denmark

10 Mar 2023
14:30-16:00

The need for FAIR data is urgent because of the growing amount of digital data, and the necessity to preserve and reuse research results. However, producing FAIR data can be difficult for researchers who are not familiar with digital methods. Furthermore, the funding authorities have focused especially in defining criteria for the assessment of FAIR repositories.

We describe some of the processes for supporting the production and distribution of FAIR data implemented at the European CLARIN infrastructure[1] and its Danish part CLARIN-DK,[2] and we discuss some of the obstacles which we have become aware of and are addressing in different ways.

CLARIN is the European digital infrastructure for language as social and cultural data, offering data, tools and services supporting research based on language resources. CLARIN-DK is run by the Centre for Language Technology, Department of Nordic Studies and Linguistics, at the University of Copenhagen, currently funded by the university. CLARIN-DK comprises a certified repository with Danish and multilingual resources as well as online services.

All aspects of FAIR data are addressed in CLARIN, and only data that can be shared by the public or the larger academic community can be deposited. CLARIN-DK provides Persistent Identifiers as well as a federated login system when password-protected access is needed for data with only academic use. CLARIN-DK relies on the WAYF login provided by DEIC, the Danish e-infrastructure Cooperation. Multiple initiatives at the European and national CLARIN level address common vocabularies, standards, GDPR issues, search facilities, and knowledge sharing.

Even with an infrastructure such as CLARIN supporting the production and distribution of FAIR data, we have met the following obstacles to the sharing of FAIR data:

1. Data might be a treasure that researchers do not want to share.
2. It can be difficult for researchers to see how other can use their data.
3. Data might have been created for specific purposes and, therefore, supposedly difficult to reuse.
4. Researchers mainly think of their own research and not about reusability.
5. Different research areas use different terminology.
6. It may not be easy to apply standard formats for researchers who have not worked with these in advance.
7. Lack of time.
8. Some researchers think that it is difficult to share data, and therefore they do not even try.
9. Parts in a collection can be covered by different licences/copyrights.

10. It is not always easy to determine the data licence.
11. Producing FAIR data is seldom rewarded.

The first five obstacles are mostly related to how research has been performed in many language related areas. Many researchers have worked alone and constructed specialized data, and they have had no need for sharing or reusing data. The behaviour of younger researchers is changing, and they collaborate, facing the need to produce and share standardised annotations. In recent years, CLARIN-DK has chosen to approach researchers individually, to understand their needs, explain how the production of FAIR data can be relevant for them, and which support they can get. We have implemented a workflow guiding the semi-automatic production of metadata for resources to be deposited in CLARIN-DK, and we have a curator who guides the researchers in the process and checks the data and metadata, and addresses obstacles 8–10. CLARIN-DK offers to participate in research projects, and this way provides support to the construction of relevant resources, addressing obstacles 6-7. Moreover, the knowledge centre DANSK guides smaller projects on standard and copyright issues.

However, producing FAIR data requires time and rewarding this effort should be a priority of institutions and funding agencies also in order to support new research patterns.

References

1 https://www.clarin.eu/

2 https://clarin.dk/clarindk/forside.jsp

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. B., Bourne, P. E., et al. (2016). The FAIR guiding principles for scientific data management and stewardship. Scientific Data, 3(160018)

Bahim, C., Casorrán-Amilburu C., Dekkers, M., Herczog, E., Loozen, N., Repanas K., Russell, K. and Stall, S. (2020). The FAIR Data Maturity Model: An Approach to Harmonise FAIR Assessments. Data Science Journal, 19: 41, pp. 1–7.

de Jong, F. M.G., Maegaard, B., De Smedt, K., Fišer, D. and Van Uytvanck, D- (2018). CLARIN: Towards FAIR and Responsible Data Science Using Language Resources. In Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), pp. 3259 – 3264, Istanbul, Turkey, ELRA.

Van Uytvanck, D., Stehouwer, H., and Lampen, L. (2012). Semantic metadata mapping in practice: the virtual language observatory. In Nicoletta Calzolari et al., editors, Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12), Istanbul, Turkey. ELRA, pp. 1029-1034.

Labastida Ignasi and Margoni Thomas (2020) Licensing FAIR Data for Reuse. Data Intelligence 2020, 2 (1-2), 199–207.

Community and Interoperability at the Core of Sustaining Image Archives

Ulrike Felsing[1], Peter Fornaro[2], Max Frischknecht[1,3], Julien Antoine Raemy[2,4]
[1]Bern University of Applied Sciences, Switzerland; [2]University of Basel, Switzerland; [3]University of Bern, Switzerland; [4]Swiss National Data and Service Center for the Humanities (DaSCH), Switzerland

10 Mar 2023
14:30-16:00

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 from 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.

In our paper, we discuss how the digital domain extends the physical sustainability of analogue archives through communication with the public. Our interdisciplinary research project Participatory Knowledge Practices in Analogue and Digital Image Archives (PIA) (2021–2025), aims to increase the use of image-based research data by developing participatory tools and application programming interfaces (APIs). The goal is to encourage the collaborative production of knowledge by interested communities. Data that is used is inherently more sustainable.

Our research is based on three cultural heritage collections of the Swiss Society for Folklore Studies: one focusing on scientific cartography (Atlas of Swiss Folklore, published from 1950 until 1995), a second from the estate of the photojournalist Ernst Brunner (1936–1979), and a third photographic collection owned by the Kreis Family (1860–1970).

The web-based tools we have developed allow citizen scientists to edit, enrich and curate data through a graphical user interface (GUI). On the one hand, we enable crowdsourcing in the usual sense through moderated metadata enrichment.[6,7] On the other hand, we also go beyond by enabling users to create open-ended narratives by creating their own sub-collections, annotating resources, adding perspectives by uploading their content, and collaborating with other users by launching open "Calls for images/submissions." By giving many different groups a voice in curating their collections, participation can democratise the decision-making power of archives.[8]

Collaboratively designed standards build the base by which individuals and institutions can participate by having unmitigated access to the data. We rely, in particular, on Linked Open Usable Data (LOUD) standards[9] such as specifications from the International Image Interoperability Framework (IIIF), a community-driven initiative that has developed shared APIs based on agreed-upon design principles for representing and annotating digital resources.[10] Implementing IIIF APIs into our research is a way of communicating specific values about sharing and open data practices as well as improving the resilience of cultural heritage data.

The software architecture is strictly separated into a headless backend that offers the potential to represent data in complex digital data models, various interfaces for communication, and a frontend framework that is capable of embedding tools for various types of applications. In a series of workshops and interviews with both academic and non-academic users, the project currently analyses the new demands of digital (and process-oriented) knowledge production.

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 sustainability of data and digital tools is closely related to the application; we go beyond open data by demonstrating the power of standardised APIs. The possibility to enrich data makes data sustainable and increases the attraction of digital infrastructures.

References

1 Kärberg, T., & Saarevet, K. (2016). Transforming User Knowledge into Archival Knowledge. D-Lib Magazine, 22(3/4).

2 Hinchcliffe, G., & Whitelaw, M. (2018). The Corley Explorer. State Library of Queensland.

3 Graf, N. (2022, June 23). Georeferenzierung in sMapshot. Open knowledge, what's next? Opendata.ch/2022 Forum.

4 Baggett, M., & Gibbs, R. (2014). Historypin and Pinterest for Digital Collections: Measuring the Impact of Image-Based Social Tools on Discovery and Access. Journal of Library Administration, 54(1), 11–22.

5 Carron, A. (2018). 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].

6 Ridge, M. (2016). Making digital history: The impact of digitality on public participation and scholarly practices in historical research. PhD thesis The Open University.

7 Ridge, M. (2017). Crowdsourcing our Cultural Heritage, Routledge. ISBN 9781138706170

8 Franzoni, C., & Sauermann, H. (2014). Crowd science: The organization of scientific research in open collaborative projects. Research Policy, 43(1), 1–20.

9 Sanderson, R. (2018, May 15). Shout it Out: LOUD. EuropeanaTech Conference 2018, Rotterdam, the Netherlands.

10 Snydman, S., Sanderson, R., & Cramer, T. (2015). The International Image Interoperability Framework (IIIF): A community & technology approach for web-based images. Archiving Conference, 2015, 16–21.

A Complex Philosophical Œuvre and its Complex User Community: The Case of the Wittgenstein Archives Bergen

Nivedita Gangopadhyay[1], Sebastian Sunday Grève[2], James Matthew Fielding[3], Alois Pichler[1]

[1]University of Bergen, Norway; [2]Peking University, China; [3]Takin.solutions, Bulgaria

10 Mar 2023
14:30-16:00

In this paper, we discuss the results and implications of a recent user survey we conducted for the digital resources at the Wittgenstein Archives at the University of Bergen (WAB). WAB was established in 1990 and is a research infrastructure and projects platform bringing together philosophy, editorial philology, text technology and digital humanities. WAB contains works by the philosopher Ludwig Wittgenstein (1889-1951) and is perhaps best known for the publication of Wittgenstein's Nachlass: The Bergen Electronic Edition (BEE, Oxford University Press 2000). On his death in 1951, Ludwig Wittgenstein left behind a philosophical Nachlass of some 20,000 pages. The WAB's research infrastructure includes digital and paper copies as well as transcriptions of Wittgenstein's Nachlass as it was first catalogued by G.H. von Wright in his 1969 article "The Wittgenstein Papers" as well as later additions to the catalogue. Since the publication of the Nachlass' Bergen Electronic Edition, WAB has undertaken significant developments of its digital tools. Since 2014, WAB produces a new digital facsimile of the Wittgenstein Nachlass, which is made available open access on the Wittgenstein Source site. Since 2016, WAB has enabled interactive open access to all its transcriptions of the Wittgenstein Nachlass on the interactive dynamic presentation (IDP) site. WAB offers semantic faceted search and browsing of Wittgenstein metadata on the semantic faceted search and browsing (SFB). WAB also offers advanced tools for Nachlass text search with WiTTFind, a cooperation with the Centrum für Informations- und Sprachverarbeitung an der Ludwig Maximilians Universität München.

WAB has an extensive and diverse user-base comprising researchers from disciplines such as philosophy, computational linguistics, digital humanities, philology, literary theory and criticism, graphic designing, and musicology. Recently, we conducted a user survey of the digital tools and resources available at WAB. The survey asked users to evaluate the following available digital tools - 1) Wittgenstein IDP: The Nachlass in interactive dynamic presentation. 2) Wittgenstein SFB: Wittgenstein resources by semantic faceted search and browsing. 3) Wittgenstein Source: The Bergen Nachlass Edition and other primary sources. 4) Wittgenstein WiTTFind: The Finder app for Nachlass text search. 5) Wittgenstein XML TEI: The Nachlass in XML TEI transcription. 6) Wittgenstein OWL: The Wittgenstein domain in ontology representation. The user group was mostly selected on the basis of long-term participation in the Wittgenstein research community and familiarity with WAB's digital resources. A main goal of the user survey is to strengthen user-oriented development of WAB and integrate users' feedback in the decision making processes. Given the enormous complexity of Wittgenstein's Nachlass, both from the point of view of philosophical content, diversity of materials, open-endedness of user-scenarios and from the sheer number of manuscript pages, users' participation in developing the digital resources is indispensable. Wittgenstein's works represent a classic case of a humanities oeuvre that resists a generalised, mechanical structuring by archivists

who are not familiar with the philosophical complexities, subtleties and controversies in the works. As such, one can think of WAB as a constant work-in-progress, much like Wittgenstein envisaged his philosophy to be. For the success of WAB as an outstanding knowledge base and research tool, it is critical that the knowledge base is dynamic just like the works it represents and like the diverse uses of the works. User involvement in the development and maintenance of the resources is key to achieving this goal. In our user-survey, a standout positive feedback is that users greatly appreciate the interactive nature of some of the resources where they can adopt an editorial role. In our paper we shall discuss the results of our survey for the future development of WAB along with the challenges that we face.

References

von Wright, G.H. 1969. The Wittgenstein Papers. Philosophical Review 78 (4). 483–503.

Wittgenstein's Nachlass: The Bergen Electronic Edition (BEE). 2000. Oxford: Oxford University Press.

Wittgenstein Nachlass Interactive Dynamic Presentation (IDP): http://wab.uib.no/transform/wab.php?modus=opsjoner.

Wittgenstein Semantic Faceted Search and Browsing (SFB): http://wab.uib.no/sfb/.

Wittgenstein Source: http://www.wittgensteinsource.org/.

Wittgenstein WiTTFind: http://wittfind.cis.uni-muenchen.de/.

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

Michael Azar[1], Daniel Brodén[1], Mats Fridlund[1], Michael McGuire[2]
[1]University of Gothenburg, Sweden; [2]Indiana University Bloomington, United States of America

9 Mar 2023
14:30-16:00

The paper presents the digital history project The Cultural Imaginary of 'Terrorism' (2022–2025) that explores the cultural engagement with terrorism in Sweden during the Cold War. Drawing on History of Ideas, Media History, and Language Technology (LT), the project examines the cultural meaning-making of political terror in national newspapers and cultural periodicals as well as works of nonfiction and fiction during a critical period for the formation of the international discourse on terrorism.

The concept of terrorism is not only ambiguous, but also partly a recent historical construct that emerged during the Cold War (Stampnitzky 2013). To explore the 'cultural imaginary of terrorism' – figures thought, frames of references and fantasies – is to study how a certain culture makes meaning of terrorism and itself in relation to the phenomenon (Frank 2017). We examine the cultural imaginary of terrorism in Sweden as a site of both discursive convergence and conflict. Critical attention is paid to the extent to which terrorism has been framed as a 'foreign' or a 'domestic' issue, as well as in relation to the tension between East-West, Left-Right and North-South.

The project's 'digital strand' draws on the major national newspapers stored at the National Library of Sweden and KBLab. A range of LT approaches allows us to identify news reports and editorials related to the topic of terrorism and to create overviews and visualizations (statistics, timelines, etc.) of the terrorism discourse and the development over time of key terms ('terrorism', 'urban guerilla', 'state terrorism', etc). We may also seek out individuals, places and groups (named-entity recognition); quantitatively map terrorism-related collocations and underlying themes (topic models); identify changes in conceptual meanings and associations over time and across different publications (word vectors) (Yao et al. 2017 Underwood 2019).

The paper presents the project's digital approach through an explorative analysis of Svensk Tidskrift ('Swedish Periodical,' 1945–1991), a conservative periodical tied to the Right party. Svensk Tidskrift is publicly available in digital form (Svensk tidskrift) and will be scraped into a corpus. Using named entity recognition, the corpus is processed to extract significant authors, actors, works and events, etc. that comprised elements of the Swedish cultural imaginary of terrorism during the Cold War. Combining these named entities with key term searches and word vector similarity, we analyse socio-conceptual discourse networks composed of people, events and the written word. Drawing upon Actor-Network Theory and Controversy Mapping (Venturini & Munk 2022), we and examine what discursive elements (incidents, groups, countries, ideologies, etc.) were integrated in the framing of terrorism, paying attention to connections between different and conflicting actors and actants.

References

Frank, M (2017): The cultural imaginary of terrorism in public discourse, literature and film, Routledge.

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

Underwood, T (2019): Distant horizons, University of Chicago press.

Venturini, T & A K Munk (2022), Controversy mapping, Wiley.

Yao, Z, Y Sun, W Ding, N Rao & H Xiong (2017): 'Dynamic word embeddings for evolving semantic discovery', International conference on web search and data mining, WSDM.

Detection and Clustering of Printers' Marks to Reveal the Publisher Networks of 18th Century Books

Ruilin Wang, Yann Ryan, Lidia Pivovarova, Mikko Tolonen
University of Helsinki, Finland

9 Mar 2023
12:30-14:00

Eighteenth-century publishing was a transitional phase in book production. The market grew, and the printing business was based on new modes of collaboration. Whereas in earlier times the printer was the main entity behind the physical production of books, the eighteenth century saw the emergence of publishers, which played multiple roles. Many individuals mixed all of these roles by owning a printing shop, buying in copyright shares and taking care of book distribution. Besides archival sources, the place to find this information is the title page. However, even in the best bibliographic records we have available, it is often missing, particularly in the case of the printer.

One promising method for recovering missing printer information is through the use of visual information. Printers used a wide variety of 'ornaments': visual elements such as 'devices', decorative initials, fleurons, and headpieces. Taken together, these can form valuable clues as to their identities. In this paper, we describe a method by which we detect and cluster one category of ornament—the decorative initial—and apply it to the problem of missing printer information.

Previously, Wilkinson et. al. (2021) applied traditional computer vision techniques for printmark detections on the whole ECCO dataset, but they did not classify the symbols into specific categories. Other projects have focused on the detection of decorative initials, or 'lettrines' (Uttama et. al., 2006; Nguyen et. al., 2020), or have used visual elements to aid in the discovery of document provenance (Hu et al., 2015), as well as making illustrations available for visual analysis (Dutta et al., 2021).

First, we manually annotated 6,322 visual elements on a random sample of 7,000 pages taken from a large dataset of digitised books (ECCO) into 15 subcategories of 5 main categories. With this training data, we implemented two deep learning methods: MaskRCNN and EfficientDet, to detect and categorise visual elements and then compared their performance with Fleuron, the model Wilkinson proposed in 2021. EfficientDet and MaskRCNN are efficient and scalable deep learning methods commonly used in object detection (Anantharaman et. al., 2018). The results were evaluated using the Mean Average Precision (MAP) method using a particular threshold of the intersection over union(IoU) score (75% in our case). In our evaluation, MaskRCNN outperformed the EfficientDet and Fleuron models in both main category and subcategory detection tasks with MAP75 scores of 64.0% and 60.9%.

In a further step, we created more training data specifically for the initials. Likely, candidates for initials were detected by applying the existing model to the first five pages of a subset of ECCO, detecting 13,425 initials from 24,886 pages. From this, we subsetted all pages on which an initial was detected with a confidence score of 10%–40%, resulting 3,217 suspicious pages and carried out a further set of manual annotations on a random subset of 1,000 pages. This improved our model results for initials from 83.30% to 98.86% in MAP75.

We then applied this initial-detection model to books published by members of an important eighteenth-century publishing dynasty: the Tonson family. Out of 2,494 books published (as recorded in the ESTC), we detected 9,202 initials with a confidence score of over 90%, spread across 9,154 pages. For these initials, we removed the final model layer and used the resulting embeddings as features for a number of clustering and dimension-reduction algorithms. We then compared the overall initial similarities of books with unknown printers to a set with the known printer information. This information is key to understanding more about the networks of publishers and printers which produced eighteenth-century texts.

References

Anantharaman, R., Velazquez, M., & Lee, Y. (2018). Utilizing Mask r-cnn for Detection and Segmentation of Oral Diseases. 2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), 2197–2204.

Dutta, A., Bergel, G., & Zisserman, A. (2021). Visual Analysis of Chapbooks Printed in Scotland. The 6th International Workshop on Historical Document Imaging and Processing, 67–72.

Hu, B., Rakthanmanon, T., Campana, B. J. L., Mueen, A., & Keogh, E. (2015). Establishing the Provenance of Historical Manuscripts with a Novel Distance Measure. Pattern Analysis and Applications, 18(2), 313–331.

Nguyen, N.-V., Coustaty, M., & Ogier, J.-M. (2020). An Adaptive Document Recognition System for Lettrines. International Journal on Document Analysis and Recognition (IJDAR), 23(2), 115–128.

Uttama, S., Loonis, P., Delalandre, M., & Ogier, J.-M. (2006). Segmentation and Retrieval of Ancient Graphic Documents. In W. Liu & J. Lladós (Eds.), Graphics Recognition. Ten Years Review and Future Perspectives (pp. 88–98). Springer.

Wilkinson, H., Briggs, J., & Gorissen, D. (2021). Computer Vision and the Creation of a Database of Printers' Ornaments. Digital Humanities Quarterly, 015(1).

Developing a Thesaurus-Based Semantic Tagger for Danish

Ross Deans Kristensen-McLachlan[1], Nicole Dwenger[1], Sanni Nimb[2]
[1]Aarhus University, Denmark; [2]Society for Danish Language and Literature, Denmark

8 Mar 2023
14:30-16:00

Semantic analysis of natural language is a fundamentally important task for a wide range of research disciplines. Many of the most successful current approaches to this problem draw on unsupervised machine learning methods which are essentially language agnostic. Despite substantial advances made in this area in recent years, these approaches still have limitations. For example, the relevant learning algorithms are both data-hungry and computationally intensive. Moreover, the sophisticated mathematical nature of these models often means that they are conceptually inaccessible to researchers in the humanities. This means that something like fine-tuning contextual embeddings on a specific domain can turn out to be a fraught and complex task. This poses a serious challenge to researchers who might want to use computational tools in order to better understand their data. These limitations might be more practical than theoretical, but they are limitations nevertheless.

An alternative approach to this problem would be to draw on existing knowledge of the lexical semantics of the language in question, leveraging existing resources to assign semantic annotations to linguistic data (Paio et al. 2017). While less computationally sophisticated, such an approach in fact has a number of benefits. Firstly, the outputs are more immediately understandable for scholars already working with close reading of text and language data, narrowing the divide between quantitative and qualitative analyses. Secondly, it minimises computational overheads, making deep semantic analysis more accessible for researchers without access to advanced computing skills and hardware while also reducing environmental impact. Finally, it is less linguistically naïve, insofar as it acknowledges and incorporates linguistic structure into the process of tagging texts. Interpretability and usability is prioritised over perceived conceptual and technical sophistication.

This approach necessarily requires a language to have a rich set of lexical resources which can be drawn on when creating a tagger. Danish is one such language, with Den Danske Ordbog, Ordbog over det danske Sprog, and Den Danske Begrebsordbog. The digitisation of these resources means that we are now in an ideal position to utilise years of detailed linguistic research which they contain. This paper presents a prototype thesaurus-based tagger for semantic annotation of Danish language texts. To do this, it draws on the conceptual structure found in Den Danske Begrebsordbog (Nimb et al. 2015) as a way of capturing hierarchical semantic relations in the Danish language. This allows for analysis at different levels of the semantic hierarchy - from more fine-grained nuances up to broader conceptual categories.

At present, no gold standard dataset exists against which to evaluate the performance of such a tagger. Evaluation was therefore conducted by having three human annotators evaluating the output from the tagger. Focusing specifically on nouns, annotators agreed that the tagger returns the correct semantic category in around 70% of cases, with a reasonable degree of agreement between annotators (F1=0.67; k=0.45). While there is still work to do, this work suggests that a conceptually and computationally simple algorithm combined with rich lexicographical resources can still offer. While the data from Den Danske Begrebsordbog is copyrighted, the

tagger itself is open source and accessible via the project Github repo. The system itself could be easily adapted to work with other languages with similar linguistic resources, such as many of those represented at DHNB.

References

S. Piao, F. Dallachy, A. Baron, J. Demmen, S. Wattam, P. Durkin, J. McCracken, P. Rayson, M. Alexander, A time-sensitive historical thesaurus-based semantic tagger for deep semantic annotation, Computer Speech & Language 46 (2017) 113–135.

R. D. Kristensen-McLachlan, N. Dwenger, Tagging Danish texts using Den Danske Begrebsordbog, 2022.

S. Nimb, L. Trap-Jensen, H. Lorentzen, L. Theilgaard, T. Troelsgaard, Den Danske Begrebsordbog, Det Danske Sprog- og Litteraturselskab, 2015

The Diachrony of the New Political Terrorism: Tracing Neologisms and Frequencies of Terror-related Terms in Swedish Parliamentary Data 1971–2018

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

University of Gothenburg, Sweden

9 Mar 2023
14:30-16:00

The paper contributes to the methodological innovation in the research on the discourse on terrorism through a data-rich analysis of the uses of the concept in Swedish parliamentary debate, drawing on an extensive corpus of debate transcripts from the Swedish unicameral Parliament (provided by the WESTAC project and the Riksdag's Open Data). Combining the affordances of language technology with domain expertise in digital history, political science, linguistics, and terrorism studies, we provide a systematic account of the conceptual development of the contemporary discourse of terrorism in Swedish parliamentary deliberations concerning terrorism, focusing on conceptual productivity and diachronicity (see Säily, Mäkelä & Hämäläinen 2021; Jarlbrink et al. 2022).

The study departs from the scholarly view that the concept of terrorism is not only pejorative and ambiguous, but also partly a recent historical construct (Stampnitzky 2013). It extends a previous study (Fridlund et al., 2022) on the conceptual productivity of the sibling concepts 'terror' and 'terrorism', i.e. the introduction of neologisms and new word forms ('terror bombing', 'terrorism deed', 'union terror', etc.) of the Swedish bicameral Parliament 1867–1970. As our prior investigation found that 'terrorism' started to be used in a contemporary sense in Swedish parliamentary debate in the late 1960s, this study will focus on the further development of this concept and its word forms from 1971 up until the late 2010s, a period that marked the rise of the modern (Western) discourse of terrorism.

Key research questions are: To what extent has the concept of 'terrorism' developed neologisms, as manifested in simplexes, compounds and collocations, and in what contexts have those new word forms have appeared? To what extent have party affiliations and the left-right conflict dimension in Swedish politics factored into the productivity of highly specific and value-laden neologisms?

The paper will chronologically trace the introduction of words that have become key reference points in contemporary terrorism discourse – 'left-wing terrorism', 'right-wing terrorism', 'Islamist terrorism', 'terrorism journey', etc. – and the frequency of their use in the parliamentary debate in Sweden.

Combining distant and close reading, we present the results in the form of chronological and frequency graphs and interpretative analysis of the different word forms of terrorism during the period in focus. This enables us to explore the discursive contexts of the multitude of word forms related to the concept of 'terrorism' in a historically nuanced way. The paper will touch on parliamentary debates concerning a plethora of historical issues, including the killing of the Yugoslavian ambassador in Sweden by Croatian separatists in 1971, the occupation of the West German embassy in Stockholm by RAF militants in 1975, the averted kidnapping plan against former minister Anna-Greta Leijon in 1977, the enforcement of the counter-terrorism legislation against Kurdish militants associated with the PKK in the 1980s as well

as the rise of right-wing and Islamist terrorism since the 1990s, including the truck attack on Drottninggatan in downtown Stockholm 2017.

References

Säily, Mäkelä & Hämäläinen (2021): 'From plenipotentiary to puddingless: Users and uses of new words in early English letters', M Hämäläinen, N Partanen, K Alnajjar (eds): Multilingual facilitation, University of Helsinki.

Fridlund, M, D Brodén, V Wåhlstrand Skärström (2022): 'The diachrony of political terror: Tracing terror and terrorism in Swedish parliamentary data 1867–1970'.

E Volodina et al. (eds): Live and learn: Festschrift in honor of Lars Borin, Gothenburg: Research Reports from the Department of Swedish, Multilingualism, Language Technology.

Jarlbrink, J F Norén & R Saberi (2022): 'Contextual modelling of "propaganda", "information" and "upplysning" in Swedish Parliamentary Speeches, 1920–2019', DiPaDA 2022, Uppsala University.

Stampnitzky, L (2013): Discipling terror: How experts invented 'terrorism', Cambridge univ. press.

The Digital Lab as an Arena for Teaching and Outreach Activities Connected to the Special Collections at the University of Bergen Library

Emma Josefin Ölander Aadland
University of Bergen, Norway

10 Mar 2023
12:30-14:00

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, discuss, and apply various digital tools and methods used in research within the Digital Humanities. Each semester the Digital 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 activities connected to the exhibitions at the Arts and Humanities Library, where different parts of the special collections at the library are displayed. The aim is to explore how these activities can have a cross-disciplinary and collaborative approach, and to give some suggestions to answer the following question:

What kind of challenges and possibilities may DH-activities connected to the exhibitions in the library provide the Digital Lab for community-building and sustainable operation?

The hypothesis is that the Digital Lab can play a part in already existing academic activities both at the library and in different academic communities and use this as a sustainable way of operating as a cross-disciplinary hub for students, researchers, and lecturers. At the same time, the lab may figure as an arena for the special collections concerning outreach activities, which can contribute to making the collections more available and involve them as subjects for collaborative activities such as DH-projects or research.

The special collections at the library consist of the Language Collections, the Picture Collection, the Rare Book collection, and the Queer archive. The first exhibition in the Arts and Humanities library was set up in 2021 after an upgrading and renewing process that gave a space aimed for exhibitions in the library.

The Digital Lab has developed a concept where at least two contributors that have worked with the different exhibitions (both physically and digitally) directly (for example, as a curator or professional advisor) or as a researcher from different fields are invited to the lab for a seminar. The concept is to give the audience an insight into how the contributors have worked with digital (or digitized) material in connection with the exhibition and/or in their own research.

References

Bell, E. C. & Kennan, M. A. (2021). Partnering in Knowledge Production: Roles for Librarians in the Digital Humanities, Journal of the Australian Library and Information Association, 70(2), 157-176.

Burns, J. A. (2016). 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(2-3), 238-248.

Caswell, ML. (2021). "'The Archive' Is Not an Archives: On Acknowledging the Intellectual Contributions of Archival Studies". UCLA.

Deegan, M., McCarty, W., & Short, H. (2012). Collaborative research in the digital humanities: a volume in honour of Harold Short, on the occasion of his 65th birthday and his retirement, September 2010. Farnham: Ashgate.

Fay, E. & Nyhan, J. (2015), Webbs on the Web: libraries, digital humanities and collaboration, Library Review, (64:1-2), 118-134.

Gooding, P. (2020) The library in digital humanities: interdisciplinary approaches to digital materials. In: Schuster, K. and Dunn, S. (eds.), Routledge Handbook on Research Methods in Digital Humanities. Series: Routledge international handbooks (pp. 137-151). Routledge: Oxon, UK.

Hartsell-Gundy, A., Braunstein, L. & Golomb, L. (2015). Digital humanities in the library: challenges and opportunities for subject specialists. Association of College and Research Libraries.

Kear, R. & Joranson, K. (2018). Digital humanities, libraries, and partnerships: a critical examination of labor, networks, and community. Chandos Publishing.

Millson-Martula, C. & Gunn, K. (2017). The digital humanities: Implications for librarians, libraries, and librarianship, College & Undergraduate Libraries, 24(2-4), 135-139.

Mapes, K. (2020). Discovering Digital Humanities Methods Through Pedagogy. In: Schuster, K. and Dunn, S. (eds.), Routledge Handbook on Research Methods in Digital Humanities (pp. 331-35). Series: Routledge international handbooks. Routledge: Oxon, UK.

Nichols, J., Melo, M., & Dewland, J. (2017). Unifying space and service for makers, entrepreneurs, and digital scholars. Portal: Libraries and the Academy, 17(2), 363-374.

Smithies, J. & Ciula, A. (2020). Humans in the loop. Epistemology and method in King's Digital Lab. In: Schuster, K. and Dunn, S. (eds.), Routledge Handbook on Research Methods in Digital Humanities. (pp. 155-172). Series: Routledge international handbooks. Routledge: Oxon, UK.

Spiro, L. (2012). "This is why we fight": Defining the values of the digital humanities. In M. K. Gold and L. F. Klein (eds.), Debates in the Digital Humanities (pp. 16-35). University of Minnesota Press.

Svensson, P. (2016). Introducing the Digital Humanities. In Big Digital Humanities: Imagining a Meeting Place for the Humanities and the Digital (pp. 1-35). University of Michigan Press.

Sula, C. A. (2013). Digital Humanities and Libraries: A Conceptual Model, Journal of Library Administration, 53 (1), 10-26.

Sula, C. A., Hackney, S. E., and Cunningham, P. (2017). A survey of digital humanities programs. The Journal of Interactive Technology and Pedagogy (11). A Survey of Digital Humanities Programs / (cuny.edu)

Poremski, M. D. (2017). Evaluating the landscape of digital humanities librarianship, College & Undergraduate Libraries, 24(2-4), 140-154.

Posner. (2013). No Half Measures: Overcoming Common Challenges to Doing Digital Humanities in the Library. Journal of Library Administration, 53(1), 43-52.

Vandegrift, M. (2012). What Is Digital Humanities and What's it Doing in the Library? In the Library with the Lead Pipe, 2012. What Is Digital Humanities and What's it Doing in the Library? – In the Library with the Lead Pipe

Wilms, L., Derven, C., O'Dwyer, L., Lingstadt, K., Verbeke, D., & Lefferts, M. (2019). Europe's Digital Humanities Landscape: A Study From LIBER's Digital Humanities & Digital Cultural Heritage Working Group. LIBER Europe.

Digital Witnesses. Representation of the Bucza Genocide on the Social Media Platforms

Bartosz Hamarowski, Maria Lompe
Nicolaus Copernicus University, Poland

9 Mar 2023
14:30-16:00

On March 29, 2022 – more than a month after the beginning of Russia's armed aggression against Ukraine – the world was flooded with information about the genocide committed in Bucza by Russian forces. Alongside official messages published by international and local journalistic outlets, photographs documenting the genocide were released primarily on social media. Many of them were considered by Facebook and Instagram algorithms as the so-called "sensitive content", which entailed blocking access, deleting or "blurring" images of the Bucza massacre.

In light of these events, we pose the question of the role played by social media in the coverage of genocide. Taking after Richard Rogers (2009) in assuming no distinction between online and offline, we seek to show the hybrid connections between these spheres. Drawing from research in the field of visual analysis of various political and social movements and events, we conceptualize visual representations as important digital artefacts that shape collective consciousness (Niedererer and Colombo 2019; Marres, Suárez Val, Tripp et al. 2020).

During the conference, we will present the results of a study of images published between March 29 and September 29, 2022, on social media platforms (Twitter, Facebook, Instagram). These include the findings from a temporal and cross-platform analysis using digital and computational methods. The 4CAT and CrowdTangle tools were used for data extraction. The raw data set consisted of a total of 74498 images (Facebook: 4811, Twitter: 68240, Instagram: 1447). Given the vast disparity between the number of images from Twitter in comparison to other platforms, the following stages of the analysis included only the 10% most retweeted images, as well as an equivalent number of images obtained through the random sampling method. The data was cleaned from mismatched, non-subject-related visual representations, which were derived from search query ambiguity and users' accidental tagging. The exploratory analysis of visual representations was performed using software for analyzing large image collections (ImageSorter, ImagePlot). This was followed by manual verification of the validity of the algorithm's clustering.
The accomplished research objectives include:

(a) assessment of the level of interest over time measured by the number of entries;

(b) delineation of the types of visual representations and their prevalence over time;

(c) description of how representations of the Bucza massacre vary across social media platforms;

(d) characterization of the policies of each platform and their impact on the visual representation of the Bucza genocide;

(e) assessment of the degree to which social media platforms influence the representation of the Bucza genocide.

With the prospect of a troubled future looming, we believe that our paper can become an important voice in the upcoming academic and political debates on the role of social media in documenting and witnessing war crimes.

References

Ciuriak, D. (2022). The Role of Social Media in Russia's War on Ukraine.

CrowdTangle Team (2022). CrowdTangle. Facebook, Menlo Park, California, United States.

Davidjants, J., Tiidenberg, K. (2022). Activist memory narration on social media: Armenian genocide on Instagram. New Media & Society, 24(10), 2191–2206.

Garner, I. (2022). "We've Got to Kill Them": Responses to Bucha on Russian Social Media Groups, Journal of Genocide Research.

ImageSorter 4.3 (2022). Available at https://imagesorter.software.informer.com/.

Marres, Suárez Val, Tripp et al. (2020). Corona Testing on Twitter. Surfacing testing situations beyond the laboratory, Digital Methods Initiative.

Niederer, S. & Colombo, G. (2019). Visual Methodologies for Networked Images: Designing Visualizations for Collaborative Research, Cross-platform Analysis, and Public Participation. Diseña, (14), 40-67.

Pearce, W., Ozkula SM., Greene AK., et al. (2020). Visual cross-platform analysis: Digital methods to research social media images. Information, Communication & Society 23(2): 161–180.

Peeters, S., & Hagen, S. (2022). The 4CAT Capture and Analysis Toolkit: A Modular Tool for Transparent and Traceable Social Media Research. Computational Communication Research 4(2).

Rogers, R. (2009). The End of the Virtual – Digital Methods, Amsterdam University Press.

Rogers, R. (2021). Visual media analysis for Instagram and other online platforms. Big Data & Society, 8(1).

Software Studies Initiative (2011). ImagePlot software v. 1.1.
Available at: http://lab.softwarestudies.com/p/imageplot.html (accessed 10th October 2022).

Distant Reading the Climate: Digital Analysis of Weather Information in 19th Century Press

Krister Kruusmaa
National Library of Estonia, Estonia

8 Mar 2023
14:30-16:00

From descriptions of calamities to mundane reports on wind speed, news about the weather has been a continuous element in the press, even long before daily forecasting. Despite being valuable testimonies about past societies' relationship to the environment, the press has been a relatively underexploited resource in the field of climate history. This paper examines the presence of weather-related information in a large corpus of digitized newspapers – ca. 300 000 articles of the German-language newspaper Rigasche Zeitung published in Riga during the 19th century.[1] The main objective is to gain new insights into the representation of the climate in the public written discourse. Another aim is exploring the potential of digital methods – an ingredient which, although not uncommon in climate history, is rather rarely used in the strenuous phase of data collection.

The first part of this paper looks at mentions of weather phenomena on the level of isolated keywords. There are several challenges when trying to quantify the mentions of a climate-related word, such as Sturm (storm), for example. Above all, the amount of OCR noise and the particularities of historical text make it impossible to achieve meaningful results with traditional NLP techniques. On the other hand, a fixed set of keywords would not work either, as there is an undetermined number of surface forms for any given word in the corpus. I present a novel approach to circumventing this problem, which relies on the application of vector similarity to converge different wordforms into standardized ones, and also allows to gather (historical) synonyms not known beforehand to the researcher.

The second part is dedicated to the contexts surrounding individual climate-related words. Using a recent topic modelling method, top2vec,[2] I model the segments of text that contain mentions of weather phenomena. Topic analysis helps to systematize the 19th-century weather discourse and observe its evolution over time. The hierarchical representation provided by top2vec allows to divide topics into smaller subtopics, giving a more fine-grained insight where needed. Topic modelling also solves the problem presented by the metaphorical usage of certain weather-related words (think a storm in a teacup, the wind of change, etc. for English). As such contexts differ from actual reports on the weather, some topics can be considered false positives.

The results offer a multidimensional picture of the weather as it was perceived by the 19th century reading public in the Baltic provinces. Quantifying the mentions of different weather phenomena, both local and exotic, informs us about society's relationship to the elements and allows the climate historian to scour his sources much more effectively. Particularly interesting are the results of topic modelling. It seems that the depiction of weather steadily moved from detailed descriptions of extreme events (hurricanes, shipwrecks, hailstorms) to a more routine and numerical reporting, apparently driven by the advent of telegraph communications and railway transport. The focus shifted from outliers to regularities, effectively 'making the weather quotidian', as expressed by Jan Golinski.[3] The findings seem to reinforce

the theses of Jean-Baptiste Fressoz and Fabien Locher, who argue that modernity brought about a desensitization to the climate and a disregard for the human ability to influence it, eventually leading to the 20th-century blindness to climate change.[4]

References

1 Project GitHub: https://github.com/krkryger/clim-dist

2 Angelov, D. (2020). Top2Vec: Distributed Representations of Topics (arXiv:2008.09470). arXiv.

3 Golinski, J. (2003). Time, Talk, and the Weather in Eighteenth-Century Britain. Weather, Climate, Culture (pp. 17-38). S. Strauss, B. Orlove (ed.). Oxford & New York: Berg.

4 Fressoz, J.-B., Locher, F. (2020). Les Révoltes du ciel : Une histoire du changement climatique XVe-XXe siècle. Paris: Seuil.

Plath, U., Raudkivi, P., Vanamölder, K., Kruusmaa, K., Liiv, A. H. "Kuidas kodeerida kliimat? Eesti ajaloolise kliimauurimise digitaalsest pöördest" = "How to encode the climate? On the digital turn of Estonian climate research", Keel ja Kirjandus, 64–8 (sept. 2021), p. 819-840.

Embed, Detect and Describe (EDDe): A Framework for Examining Events in Complex Sociocultural and Historical Data

Melvin Wevers[1], Jan Kostkan[2], Kristoffer Laigaard Nielbo[2]
[1]University of Amsterdam, The Netherlands; [2]Aarhus University, Denmark

8 Mar 2023
13:00-14:00

Introduction

While humanities often discuss events theoretically, there is a disconnect between theoretical work on events and empirical studies of events.[1] This paper introduces the Embed, Detect, and Describe (EDDe) framework, that is, an information-theoretical approach to (historical) event detection and characterization in noisy and complex sociocultural data. EDDe is based on the fundamental embedding theorem[2,3,4] which allows us to approximate the dynamics of a large-scale social system. Rather than measuring cultural expressions through word counts over time, we approach society as a complex system with a multitude of states which switch between attractors. Some of these attractors may be associated with the dynamics of cultural information and captured in low-dimensional indicator variables.[5] In our case, these simple indicator variables are expressed through surprise in the textual content, but EDDe is data- and medium-agnostic.

Event Detection

To detect state changes, or events, in a complex system, it is necessary to extract reliable information about states and transitions in the system. We use an information-theoretic approach to extract simple indicator variables from large collections of cultural data by modelling information states with windowed relative entropy between dense low-dimensional text representations. EDDe uses a Bayesian approach to change detection that identifies relevant change points. This approach to change detection locates reliable shifts in the central moments of a time series. We start with a time series with non-homogeneous regions in terms of the series' mean and variance. State changes are then modelled as several transitions between locally stable means and variances. Finally, it is possible to identify locations of change and extract homogeneous regions.

Event Characterization

EDDe can be applied to event characterization, that is, how events have affected the information states in complex systems. Events can, for example, disrupt the news flow by decreasing the amount of novel information presented in media. In the run-up to an event, an increasing focus of the public's eye might be reflected in the increasing uniformity of media discourse. EDDe applies Dynamic-Time Warping Barycenter Averaging (DBA) and hierarchical clustering to the time series surrounding selected events. This yields a taxonomy showing five different ways in which events impacted the flow of the news over time. These clusters give us a better understanding of the relationship between events and their impact on news sources.

Concluding Remarks

In this paper, we present two methods for event detection and characterization. These methods are fundamentally language-independent and data-agnostic, making them suitable for wide applications of comparative analysis across temporal and geographical dimensions. Rather than examining stationary signals presented in simple line graphs of word counts, we approach cultural systems in their complexity, drawing upon theoretical and methodological work in complexity science and chaos theory. This allows us to approximate the dynamics of a large-scale social system. In the full paper, we have also added validation methods that confirm that the semantic dynamics do indeed reflect real-world events.

References

1 Theo Jung and Anna Karla. 1. Times of the Event: An Introduction. 60(1):75–85.

2 N. H. Packard, J. P. Crutchfield, J. D. Farmer, and R. S. Shaw. Geometry from a time series. Phys. Rev. Lett., 45(9):712–716, 1980. Publisher: American Physical Society.

3 Floris Takens. Detecting strange attractors in turbulence. In David Rand and Lai-Sang Young, editors, Dynamical Systems and Turbulence, Warwick 1980, pages 366–381. Springer Berlin Heidelberg, 1981.

4 Tim Sauer, James A. Yorke, and Martin Casdagli. Embedology. Journal of Statistical Physics, 65(3):579–616, 1991.

5 Jianbo Gao, Yinhe Cao, Wen-wen Tung, and Jing Hu. Multiscale Analysis of Complex Time Series: Integration of Chaos and Random Fractal Theory, and Beyond. Wiley-Interscience, 2007.

Engineering Terrorismmindedness: A Scientometric Study of the 9/11-effect on STEM Research, 1989-2013

Mats Fridlund[1], Gustaf Nelhans[2]

[1]University of Gothenburg, Sweden, [2]University of Borås, Sweden

9 Mar 2023
14:30-16:00

'9/11 changed everything', is a common phrase emphasizing there is a new 'before' and 'after' in all areas of society caused by the September 11, 2001, terrorist attacks in the USA. Following the Cold War, we have experienced a new war in the US-led 'Global War on Terrorism', and this study investigates how terrorism has shaped science and technology through the establishment within engineering research of a 'terrorismmindedness' (Fridlund 2011), i.e. how the terrorist threat becomes domesticated and normalized by being integrated in research practice. This extends research within the history of science and technology, focusing on the impact on research in science, technology, engineering and medicine (STEM) of wars in general and the Cold War in particular (see MacKenzie 1990, Leslie 1993, Hecht 1998, Abbate 1999).

Few historical studies exist of terrorism's impact on STEM (for exceptions, see Reppy 2008, Moreno 2012). Towards such a history, we provide a distant reading investigation of how research in the engineering sciences during 1989-2013 shifted towards increasingly addressing terrorism-related topics. In this, we extend earlier research (Fridlund & Nelhans 2011a, 2011b, 2011c) that demonstrated a 9/11-effect on STEM-research 2001-2010 but where we did not conduct any detailed and extensive analysis.

This study consists of an in-depth bibliometric analysis of scientific articles having the term *terroris* within its title, abstract or keywords published 1989–2013 within the 'Engineering' subject area identified in the Thomson Reuters Science Citation Index and the Conference Proceedings Citation Index (Web of Science – WoS), claimed to be representing the 'most relevant' international scientific research publications. Thus we define the imagined but real research field 'Terrorism Related Engineering Research' (TRER). Our preliminary research have identified almost 2.000 TRER articles in addition to a broader STEM-set of about 10.000 articles, retrieved for baseline reference within all WoS Science-related databases.

These TRER publications was bibliometrically mapped according to topical properties (e.g. bibliographic coupling at the journal level), clustering together research citing similar sources and using common concepts. More specifically, one of our methods identify relevant TRER topics using Algorithmic Historiography through the HistCite tool (Garfield et al., 2003) that visualize the citation network as a tree structure or family tree to reconstruct the articles' history of intellectual influences. Additionally, topicality is studied by looking at co-culture measures, making it possible to find terrorism-related research without terrorism-terms. Papers cited together are mapped using the VOSviewer software by analyses of the publications' contents (van Eck and Waltman, 2010) that not only connects terms found in the texts, but also help by identifying concepts based on noun-phrases. The results are clustered according to topicality, revealing different core TRER research interests (such as building blast-protection or bioterrorism mitigation). Additionally, text-based techniques, based on co-word analysis, together with algorithms for the

extraction of methods, theories and activities mentioned in the article sets using question-answering approaches and names, locations and entities based on named entity recognition (NER) are used to algorithmically "read" the texts and provide insights into the scientific work done in the publications. The visualizations will be investigated quantitatively and qualitatively, where identified key publications in the clusters will be closely read for qualitative indications of 9/11-effects on the research.

The resulting co-word maps are means of distance-reading the thousands of texts in the sets. Such indicator-based methods provide insights into the literature at an abstract level practically impossible otherwise. Furthermore, although these methods are highly quantitative, they have a clear qualitative stance. The clusters found are not automatically given but depend on choices regarding cut off points that lead to variations in the number of mapped clusters and the level of detail in the analysis.

References

Mats Fridlund, Daniel Brodén, Tommi Jauhiainen, Leena Malkki, Leif-Jöran Olsson & Lars Borin, "Trawling and Trolling for Terrorists in the Digital Gulf of Bothnia: Cross-lingual Text Mining for the Emergence of Terrorism in Swedish and Finnish Newspapers, 1780–1926", in Darja Fišer & Andreas Witt, eds. CLARIN: The Infrastructure for Language Resources, Digital Linguistics Series 1 (Berlin: De Gruyter, 2022), 781–802.

Magnus P. Ängsal, Daniel Brodén, Mats Fridlund, Leif-Jöran Olsson, & Patrik Öhberg, "Linguistic Framing of Political Terror: Distant and Close Readings of the Discourse on Terrorism in the Swedish Parliament 1993–2018", in Tomaž Erjavec & Maria Eskevich, eds. CLARIN Annual Conference Proceedings 2022, Series CLARIN Annual Conference Proceedings (Prague: CLARIN, 2022), 69–72.

Patrik Öhberg, Daniel Brodén, Mats Fridlund, Victor Wåhlstrand Skärström & Magnus P. Ängsal, "Unifying or Divisive Threats? Anxiety about Political Terrorism and Extremism among the Swedish Public and Parliamentarians, 1986–2020", in Karl Berglund, Matti La Mela & Inge Zwart, eds., DHNB 2022: Proceedings of the 6th Digital Humanities in the Nordic and Baltic Countries Conference (DHNB 2022), Uppsala, Sweden, March 15-18, 2022, CEUR-WS vol. 3232 (Aachen: CEUR-WS.org, 2022), 145–158.

Mats Fridlund, "Securitizing Things: Recovering a Lost Material History of the Fear of the Next War", in: Trine Villumsen Berling, Ulrik Pram Gad, Karen Lund Petersen & Ole Wæver, Translations of Security: A Framework for the Study of Unwanted Futures, Series Routledge New Security Studies 7 (Abingdon & New York: Routledge, 2022), 165–168.

Mats Fridlund, Daniel Brodén, Leif-Jöran Olsson & Magnus Ängsal, "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 Digital Parliamentary Data in Action (DiPaDA 2022) Workshop Co-located with the 6th Digital Humanities in the Nordic and Baltic Countries Conference (DHNB 2022), Uppsala, Sweden, March 15, 2022, CEUR-WS vol. 3133 (Aachen: CEUR-WS.org, 2022), 167–175.

Jens Edlund, Daniel Brodén, Mats Fridlund, Cecilia Lindhé, Leif-Jöran Olsson, Magnus Ängsal & Patrik Öhberg, "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: Proceedings of the Intelligent Systems Conference (IntelliSys) 2021, Vol. 2, Lecture Notes in Networks and Systems 295 (Cham: Springer, 2022), 435–449.

Exploring Latvian Twitter Eaters' Food-Related Sentiment in Different Weather Conditions and in Relation to Meat

Maija Kāle[1], Matīss Rikters[2]

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

10 Mar 2023
12:30-14:00

Food choice is a complex phenomenon shaped by factors like taste, ambiance, culture, weather, and many others.[1, 2, 3, 4, 5] Obesity, type 2 diabetes, and cardiovascular diseases are just a few of the health problems acquired due to the nutritional specifics of contemporary consumers. While the impact of food on personal health is an area discussed by food policymakers and nutritionists globally, another new discourse has emerged in relation to food consumption: its impact on planetary health 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, leading to meat production making up for nearly 60% of all greenhouse gases from food production.[7] Food choice and food consumption play an important role in public health, as well as impact environmental sustainability profoundly. While there is a clear policy focus towards climate-smarter diets, consumer habits remain to be slow to change.

There is insufficient knowledge of what factors influence food choice in certain days, seasons or climate. "Weather people" - this is a term that Bakhshi (2014)[8] uses to explain our dependence on the weather regarding food choices and satisfaction with food. While the weather is known to alter consumers' mood significantly and consequently their behavior, there have been surprisingly few studies that illustrate the weather impact on food perception and food choices, except some that have used online and offline restaurant reviews as a proxy of measuring it. Conclusions have been drawn that weather impacts both the frequency of the feedback that food consumers provide, as well as its content. Typically, sunny and pleasant weather leads to more frequent and more positive feedback, since low levels of humidity and high levels of sunlight are associated with high mood. At the same time, reviews written on rainy or snowy days, namely days with precipitation, tend to have lower ratings.[9]

This paper focuses on the under-researched, but influential factor that impacts food choice and perception - the weather, as well as analyses tweeters' sentiments regarding various meat containing food items. With the recent increase in the availability of datasets about food and its perception as reflected on Twitter and historical weather data, we attempt to explore food-related tweeting in different weather conditions. In this paper, we inspect a Latvian food tweet dataset spanning the past 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. Further we explore the dynamics of sentiment related to meat and meat consumption on Twitter over a ten year period.

Twitter is one of the best sources for tracing various food-related utterances in different cultures and societies since the food there is widely documented and

discussed in multiple formats.[10] We accumulate the necessary contextual knowledge of group dynamics in relation to the topic of meat consumption and consumption during different weather conditions that can be useful when aspects influencing individual food choices are examined. In this paper, we analyze social media content written in a morphological complex and less-resourced language - Latvian. This research contributes to the growing area of largescale social network data understanding of food consumers' food choices and perceptions. With a better understanding of different impact factors for food choices, we aim to contribute to public health policies that can nudge food consumers towards healthier diets and lives.

References

1 M. Kāle, J. Šķilters, M. Rikters, Tracing multisensory food experiences on Twitter, International Journal of Food Design 6 (2021) 181–212.

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

3 W. Min, S. Jiang, L. Liu, Y. Rui, R. Jain, A survey on food computing, ACM Comput. Surv.52 (2019).

4 C. Spence, Explaining seasonal patterns of food consumption, International Journal of Gastronomy and Food Science 24 (2021) 100332.

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

6 T. Lancet, We need to talk about meat, The Lancet 392 (2018) 2237, publisher: Elsevier.

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.

8 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.

9 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. arXiv.

10 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.

Exploring the Stability of Political Rhetoric in Finnish Parliamentary Debates Using Deep Learning

Otto Tarkka, Kimmo Elo, Filip Ginter, Veronika Laippala
University of Turku, Finland

9 Mar 2023
12:30-14:00

Using modern machine-learning technology to predict party affiliation based on political speeches provides new possibilities for the analysis of how political ideologies are expressed in talk. This knowledge can be applied to analyze the broader spread and development of political discourse in society. Previous work on this subject has been conducted on parliamentary speeches held in the US Congress as well at the Lithuanian, Norwegian, and Danish Parliaments (Bayram et al., 2019; Kapočiūtė–Dzikienė & Krupavičius, 2014; Navarretta & Hansen, 2020; Søyland & Lapponi, 2017). Thus far, no such work has been conducted on data from the Finnish Parliament. Our aim is to fill this gap.

In this study, we examine to what extent party affiliations can be predicted from the speeches of the Finnish Parliament and exemplify how predictive models can be used to tackle how ideologically loaded language is present in public discourses and, thus, is used to construct reality. Specifically, in this study, we explore how stable the rhetoric of political parties remains over time and seek to understand what rhetorical strategies differentiate political parties from each other. We examine, while topics of debate change, whether and to what extent the argumentation, style, and patterns of speech remain stable characteristics of each political party.

To achieve this goal, we employ deep learning. Recent developments in deep learning, especially transformer-based language models, have led to significant improvements in model performance. We use FinBERT (Virtanen et al., 2019), a Finnish transformer-based language model, and fine-tune it for classifying party affiliation from parliamentary speeches using data from plenary sessions in the Finnish Parliament. The FinParl-dataset consists of full-texts of all plenary speeches held in the Finnish eduskunta since 1907 onwards, totalling nearly a million speeches. We train several models on varying subsets of the data to explore how party affiliations are predicted in different decades and across timespans.

While deep learning models often perform better than simpler models, their complexity hinders our ability to understand them. Advances in model explainability now allow for greater insight into which features of speech deep learning models base their predictions. To understand what leads a model to make the predictions it does, we use the Python library SHAP. SHAP is a game-theoretic approach that explains how input features affect the resulting output (Lundberg & Lee, 2017). Instead of posthoc justifying model predictions based on educated guesses, we can directly inspect what textual features in a given speech contribute most to the prediction.

Finally, to get a sense of how difficult this kind of classification task is for human annotators, we conduct a human control experiment to evaluate the performance of our model. We ask a group of annotators, given a speech, to guess the party affiliation of the MP and to explain which parts of the text informed their decisions most. Using SHAP, we compare if human annotators and computer models base their predictions on the same textual features.

The results show that our models can predict party affiliation with nearly 70% accuracy in an 8 class classifying task. Accuracy drops to below 50% when a temporal

gap between train and test data is introduced. Yet, accuracy remains much higher than chance (12.5%) and majority baseline (20%), suggesting that there is some degree of stability in party rhetoric over time. Model explainability shows that models focus most on passages containing mentions of highly polarized topics (e.g., immigration, social security), party and MP names, and government-opposition contrasts.

References

Bayram, U., Pestian, J., Santel, D., Minai, A. A., & Member, S. (2019). What's in a Word? Detecting Partisan Affiliation from Word Use in Congressional Speeches 2019 International Joint Conference on Neural Networks (IJCNN).

Kapočiūtė–Dzikienė, J., & Krupavičius, A. (2014). Predicting party group from the Lithuanian parliamentary speeches. Information Technology and Control, 43(3), 321–332.

Lundberg, S. M., & Lee, S.-I. (2017). A Unified Approach to Interpreting Model Predictions. 31st Conference on Neural Information Processing Systems (NIPS 2017).

Navarretta, C., & Hansen, D. H. (2020). Identifying Parties in Manifestos and Parliament Speeches. Proceedings of ParlaCLARIN II Workshop, 51–57.

Søyland, M. G., & Lapponi, E. (2017). Party Polarization and Parliamentary Speech. ECPR 2017 General Conference.

Virtanen, A., Kanerva, J., Ilo, R., Luoma, J., Luotolahti, J., Salakoski, T., Ginter, F., & Pyysalo, S. (2019). Multilingual is not enough: BERT for Finnish.

Facilitating Data Re-use by Better Understanding Paradata Outputs

Olle Sköld, Jessica Kaiser, Lisa Andersson, Isto Huvila, Ying-Hsang Liu
Uppsala University, Sweden

10 Mar 2023
14:30-16:00

It is an important objective to achieve capable and efficient data re-use within and across many fields of research in the SSH-STEM spectrum, including the digital humanities (Martin-Rodilla and Gonzalez-Perez, 2019; Poole and Garwood, 2020). The potential commonly appended to data re-use is significant and wide-ranging: data re-use can facilitate the verification of research findings, support the exchange of knowledge, improve the efficiency of research planning and execution, and yield additional results if paired with the use of new methods or interpretative frameworks (Borgman, 2012; Whyte and Pryor, 2011). While the literature strongly shows that sufficient paradata (Couper, 2000; Sköld et al., 2022)—i.e., process data describing how data was created and curated, cf. metadata (Mayernik, 2020)—pertaining to the dataset being re-used is a key element in making data re-use practically possible and scholarly relevant (Faniel et al., 2013; Yoon, 2017), research has just begun to explore where paradata can be found and what strategies could be employed to identify and harness paradata in support of secondary data use (e.g., Börjesson et al., 2022; Huvila et al., 2021).

The present paper seeks to address this knowledge gap by elucidating what paradata outputs emerge during research, from design to enactment, reporting, and concluding data management work. Paradata outputs are identified on the basis of an interview study of researchers and professionals (n=33) working with archaeological data in different capacities. The resulting paradata-output typology is used to drive a discussion of possible strategies for capturing paradata outputs for the purpose of facilitating data re-use. The archaeological case study is valuable for understanding paradata outputs in a DH context because of the many similarities of data work in the two domains. Both DH and archaeology are characterized by data collection and processing being innovative, quickly evolving, and involving both interpretation and technological application e.g., in GIS and 3D visualization applications (Choumert-Nkolo et al., 2019; Nicolucci, 2012). The analysis of paradata outputs is guided by a genre framework (Andersen, 2008) emphasizing the close relationship between research paradata and the scholarly activity systems which they are a part of.

The findings show that paradata can be found not only in more traditional forms of documentation, but also in more informal and ephemeral contexts. At the macroscopic level, these contexts can be divided into strategic and operational groups. The strategic grouping includes examples of purposeful descriptions of data, such as data management plans, methodologies, and data dictionaries. Paradata outputs in this group were generally communicative in nature and directed toward specific audiences. The operational grouping of paradata outputs relate to the more immediate aspects of ongoing research work and include data management actions, various types of dialogue (in person, via email, or in social media), visuals, or digital signatures. Textual paradata outputs under the operational heading were often meditating in nature and primarily intended to support tasks.

The broad range of paradata outputs identified in the analysis indicates that the strategies used to capture and use paradata outputs to facilitate data re-use

in DH must be similarly broad in scope. The paper suggests that high-relevance focal points to consider in the development of such strategies are procedurality, the scholarly context, user perspectives, and paradata literacies. While the former category pertains to how paradata outputs can be harnessed – e.g., by manual or automatic means – the subsequent ones stresses the importance of having actionable insights into how researchers create paradata and what paradata data re-users require in order to be successful and the need to be able to support data re-users in gaining competencies in finding, identifying, and employing useful paradata.

References

Andersen, J. (2008). The concept of genre in information studies. Annual Review of Information Science and Technology, 42(1), 339–367.

Birnholtz, J. P. and Bietz, M. J. (2003). Data at work: supporting sharing in science and engineering. Proceedings of the 2003 International ACM SIGGROUP Conference on Supporting Group Work, 339–348.

Borgman, C. L. (2012). The conundrum of sharing research data. Journal of the American Society for Information Science and Technology, 63(6), 1059–1078.

Börjesson, L., Sköld, O., Friberg, Z., Löwenborg, D., Pálsson, G., and Huvila, I. (2022). Re-purposing excavation database content as paradata: an explorative analysis of paradata identification challenges and opportunities. KULA: Knowledge Creation, Dissemination, and Preservation Studies, 6(3), 1–18.

Choumert-Nkolo, J., Cust, H., and Taylor, C. (2019). Using paradata to collect better survey data: evidence from a household survey in Tanzania. Review of Development Economics, 23(2), 598–618.

Couper, M. P. (2000). Usability evaluation of computer-assisted survey instruments. Social Science Computer Review, 18(4), 384–396.

Faniel, I., Kansa, E., Whitcher Kansa, S., Barrera-Gomez, J. and Yakel, E. (2013). The challenges of digging data: a study of context in archaeological data reuse. Proceedings of the 13th ACM/IEEE-CS Joint Conference on Digital Libraries, 295–304.

Haslhofer, B., Isaac, A., and Simon, R. (2018). Knowledge graphs in the libraries and digital humanities domain.

Huvila, I., Sköld, O., and Börjesson, L. (2021). Documenting information making in archaeological field reports. Journal of Documentation, 77(5), 1107–1127.

Martin-Rodilla, P. and Gonzalez-Perez, C. (2019). Metainformation scenarios in digital humanities: characterization and conceptual modelling strategies. Information Systems, 84, 29–48.

Mayernik, M. S. (2020). Metadata. Knowledge Organization, 47(8), 696–713.

Niccolucci, F. (2012). Setting standards for 3D visualization of cultural heritage in Europe and beyond. In Paradata and Transparency in Virtual Heritage, eds. A. Bentkowska-Kafel, H. Denard, and D. Baker, pp. 23–36. Farnham: Ashgate.

Poole, A. H. and Garwood, D. A. (2020). Digging into data management in public-funded, international research in digital humanities. Journal of the Association for Information Science and Technology, 71(1), 84–97.

Sköld, O., Börjesson, L., & Huvila, I. (2022). Interrogating paradata. Information Research, 27.

Whyte, A. and Pryor, G. (2011). Open science in practice: researcher perspectives and participation. International Journal of Digital Curation, 6(1), 199–213.

Yoon, A. (2017). Data reusers' trust development. Journal of the Association for Information Science and Technology, 68(4), 946–956.

Finding Historical Discourse on Natural Environment: Australian Newspapers 1900-1990

Peeter Tinits

University of Tartu, Estonia

8 Mar 2023
14:30-16:00

The history of industrialized countries over the past 250 years can be understood in terms of relative neglect of the natural environment: nature has been seen as an endless source of raw materials, as something that is infinitely adaptable, capable of absorbing any pollution, and if any problems do occur, they can be easily fixed later (Kanger & Schot 2019). While a long-running trend, there are indications that a shift in popular attitudes has taken place in the past few decades. These long-term trends are however difficult to track. For more modern times, we have survey data e.g. on people's attitudes to nature conservation, but these were first collected in the 1980s only.

One development that opens up a few research avenues here is the increasing availability of digitized historical newspapers. The digitized materials open up a way to perform quantitative research on historical questions, e.g. how common were some notions in popular discourse - the sociological aspect in the history of ideas. In our case, we may ask how common was the topic of natural environment in popular discourse. We pursued this question by a keyword search - we selected a number of keywords that should get a good coverage of the contexts where nature is being discussed (nature, natural, environment* etc). These keywords are broad in scope, and should allow for high recall, however they can be very low in precision (e.g. human nature and school environment include these keywords but are not about our intended topic). To contrast this, we also chose a set of keywords that are very narrow in scope and should give results with high precision but low recall (e.g. protection of animals, nature protection, ecolog*, biodivers*).

In order to get a good set of results, we disambiguated our query results with the help of topic modelling. Topic models have been used in a number of ways to improve text retrieval and analysis for historical text collections (e.g. Oberbichler & Pflanzelter 2021, Marjanen et al. 2021, de Wildt et al. 2022). In our approach, we took the +/–25 word contexts around the keywords and built sets of topic models on them (for 10, 20, 30 topics). We annotated the topics for relevance to our initial query (e.g. topic related to animals and plants was relevant, but a topic related to school and education was not relevant). We then considered only the matches that had relevant topic content above a threshold (we tested 20% and 33%).

We used the overlap between the broad query and the narrow query to estimate the recall of the broad query. Constructing the query iteratively and disambiguating the results with topic modelling we finally found a solution that gave both high precision (67%) and high recall (78%) for our results, at the same time providing 11x more results. We analysed the frequency of the positive matches across the time period and found a significant leap in the quantity of discussion starting from around the 1960s and lasting until the 1990s, which conforms to some narratives on environmental thought (Warde et al. 2019). In the context of our study, we thus find that there are interesting changes happening in how much natural environment was being discussed. We believe that historical digitized newspapers provide a new

vantage point to look systematically at sociological questions about the past when no surveys exist. Query disambiguation and refinement via methods like topic models can become a standard part of a historians toolkit looking into these questions.

References

de Wildt, Tristan E., Ibo R. van de Poel, and Emile JL Chappin. 2022. Tracing Long-term Value Change in (Energy) Technologies: Opportunities of Probabilistic Topic Models Using Large Data Sets. Science, Technology, & Human Values 47.3: 429-458.

Laur Kanger and Johan Schot. 2019. Deep transitions: Theorizing the long-term patterns of socio-technical change. Environmental Innovation and Societal Transitions, 32:7–21.

Marjanen, Jani, Zosa, Elaine, Hengchen, Simon, Pivovarova, Lidia, & Tolonen, Mikko. 2021. Topic Modelling Discourse Dynamics in Historical Newspapers. Digital Humanities in the Nordic Countries 2020 (DHN 2020), Riga, Latvia.

Oberbichler, Sarah and Eva Pfanzelter. 2021. 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(jdh001).

Warde, Paul, Libby Robin, and Sverker Sörlin. 2018. The Environment: A History of the Idea. JHU Press, Baltimore, Maryland.

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

Maija Kāle[1], Ramesh Jain[2]

[1] University of Latvia, Latvia; [2]University of California, United States of America

10 Mar 2023
12:30-14:00

In recent years, there has been a growing interest in many areas of computer science to work with food data. One area is human-centric computing, where specific progress can be observed in integrating different senses into virtual reality, complementing it with aspects of smell, temperature, and, increasingly, taste, instead of operating only within the visual and audial formats. An increasing number of workshops are organized specifically in the context of food and human interaction, leading to the creation of a manifesto for analysing food in the context of computer science.[1] A similar development can be observed at the ACM Multimedia Conference, where the international workshop on Multimedia for cooking, eating, and related applications (CEA++) was organized in 2022, and a panel discussion "Toward Building a Global Food Network" was held.[2]

Given the challenges faced by both humans and the environment in ensuring the functioning of food systems worldwide, the development of food computing is inevitable and urgently required. The question remains: how should food computing evolve to address fundamental problems related to the food ecosystem? Could it be accomplished by deepening the scope via broadening the network, where the most crucial part is ensuring the diversity of the network and more coordinated effort with scientists from different scientific branches? Participants of CEA++'22 investigate the question "How can we create a global food network?" by tackling the challenge of data and knowledge sharing across the nations for the purpose of health and increased food systems efficiency in general.[3] Data and knowledge sharing globally is a way forward; however, we would like to rephrase the question: how can we create a diverse global food network? Diversity here implying the ability to overcome relatively narrow 'scientific discipline thinking' and invite a broader spectrum of researchers on board for future food computing development to address practical challenges to human society.

Food cultures differ from country to country and region to region, and food data produced in one area rarely works in another.[3] Awareness of this requires greater in-depth collaboration with food anthropology experts, sociologists, and food historians, who could add important knowledge to shape better research inquiries when working with food data. Community involvement in shaping understanding of various complexities related to food has proven to be successful when developing knowledge on future trends in Nordic-Baltic food systems.[4] Considering that our societies experience many inefficiencies when it comes to food (unhealthy diets, adverse planetary effects, food waste, and loss of biodiversity), while food-related big data just grows in scale and variety, Critical Reviews in Food Science and Nutrition 57 (2017) 2286–2295.[5,6] it is about time to plant and grow new, healthy, and data-driven ideas for our food and society among the research community. In other words, the time has come to diversify the scope of food computing and add interdisciplinarity and diversity to its own recipe.

We believe that for food computing the need for multidisciplinary may be broader as well as deeper in its scope and effectiveness. Food is essential for human life

and is the most important source of survival as well as enjoyment for every living person. This premise is assumed throughout this publication. We offer the future vision of food computing that will become a computing infrastructure for food. Food computing will unify all aspects of food from production to consumption and the effects of every step on people and the environment. We will build a framework for food computing that will include the food knowledge graph,[7] personal food model,[8] context for food and food recommendation engine[9] embracing the diversity that other science disciplines can offer when analysing food in its broadest sense.

References

1 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.

2 CEA++ '22: Proceedings of the 1st International Workshop on Multimedia for Cooking, Eating, and Related Applications, Association for Computing Machinery, New York, NY, USA, 2022.

3 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.

4 M. Grivins, A. Halloran, M. Kale, Eight megatrends in Nordic-Baltic food systems, Nordisk Ministerråd, 2020.

5 H. J. P. Marvin, E. M. Janssen, Y. Bouzembrak, P. J. M. Hendriksen, M. Staats, Big data in food safety: An overview

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

7 W. Min, C. Liu, S. Jiang, Towards building a food knowledge graph for the internet of food, 2021.

8 A. Rostami, V. Pandey, N. Nag, V. Wang, R. Jain, Personal food model, in: Proceedings of the 28th ACM International Conference on Multimedia, 2020, pp. 4416–4424.

9 N. Nag, V. Pandey, R. Jain, Live personalized nutrition recommendation engine, in: Proceedings of the 2nd International Workshop on Multimedia for Personal Health and Health Care, MMHealth '17, Association for Computing Machinery, New York, NY, USA, 2017, p. 61–68.

Generative Historicity: AI Image Synthesis as a Tool for Exploring Historicity and Historiography

Per Gunnar Israelson[1], Matts Lindström[2]
[1]Gävle University, Sweden; [2]Uppsala University, Sweden

9 Mar 2023
14:30-16:00

By training neural networks on huge amounts of pre-existing images and descriptive text it is possible for a computer to synthesise images purely based on textual descriptions ("prompts") provided by a human. In recent years research into such generative neural networks for image synthesis has made great leaps, especially in the field of Latent Diffusion Models. Through models like OpenAIs Dall-E and Google's Imagen, the often striking realism and artistic quality of these AI images has propelled synthetic generative art into a popular cultural phenomenon. These models however, while highly impressive, have, for ethical and commercial reasons, been closed-source and only available to the public with severe limitations. Crucially, following the public and open source release of StabilityAIs Stable Diffusion model and weights (2022), the practice of state-of-the-art LDM image synthesis is now reaching a much broader audience than before. Stable Diffusion can be freely modified, installed and runs on a consumer grade computer and GPU. This has led to an ongoing, explosive proliferation of practices and tools for image-to-text synthesis.

The increasing power and availability of such deep learning models has also spurred rich debates and important critical research concerning their cultural and ethical implications, as well as their aesthetic potential. However – given that they are based on training data sets consisting of historical images – critical questions concerning historicity and how AI image synthesis, incorporates, translates and relates to history and historiography have not yet been explored in a thorough and systematic way.

In this paper we will describe our ongoing exploration of image synthesis and "prompt engineering"–as–a–method as this relates to questions of historicity, historiography and notions of historical authenticity. Specifically, we do this while using Stable Diffusion as an exploratory device, by systematically coaxing the model into recreating historical images and objects that are part of the original training dataset – and reflecting on the difference vis-a-vis the original document. Empirically, we focus on recreating iconic images and representations picked from both popular culture (iconic photographs) and the scientific domain (scientific representations and visualisations).

The ethical and epistemological challenges accompanying representations of the past have been at the center of historiography since at least the days of Leopold von Ranke in the mid nineteenth century. The artifice of representational forms, as well as the ideological entanglements of the position of enunciation, has always conflicted with Ranke's historiographical goal of telling the past as "it really was". In the heyday of postmodernism and in the wake of the so-called linguistic turn the epistemological and ethical impact of representational forms was widely debated. With the ubiquity of computational media and the recent ecological turn in theory, the ethical and epistemological debate has come to also include the ontological status of media.

This paper resituates two concepts presented during the postmodernist debates – Linda Hutcheon's "historiographical metafiction" and Hayden White's "practical

past" – in a media theoretical discussion of synthetic image production. Adapting Hutcheon and White's concepts to a neocybernetic conceptuality the aim is to investigate the ontogenetic power of synthetic images as potential tools for generating "practical pastness".

Through this process of explorative recreation of historical representations in tandem with metahistorical reflection we gain knowledge concerning the current conditions and limits of state-of-the-art generative AI, as well as new insights into how historicity and historiography figures into and can enlighten such technological processes.

References

Elias, Amy, "Metahistorical Romance, the Historical Sublime, and Dialogic History", in Rethinking History, Vol. 9, No.2/3 2005, p 166.

Hutcheon, Linda A poetics of postmoderism: history, theory, fiction (New York: Routledge, 1988).

Serres, Michel, The Parasite, trans. Lawrence R. Schehr (Minneapolis: University of Minnesota Press, 2007).

Simondon, Gilbert (2020) Individuation in Light of Forms and Information, University of Minnesota Press.

White, Hayden, Metahistory: the historical imagination in nineteenth-century Europe (Baltimore: Johns Hopkins UP, 1973).

White, Hayden, The Practical Past (Evanston, Ill.: Northwesten University Press, 2014).

How Dark is Dark Souls? Applying Computer Vision to Analyze Video Game Walkthroughs

Thomas Schmidt, Pascal Lindemann, Maximilian Huber
University of Regensburg, Germany

9 Mar 2023
12:30-14:00

The application of computational computer vision (CV) methods has led to fascinating results and explorations in Digital Humanities (Flueckiger, 2017; Burghardt et al., 2018; Arnold & Tilton, 2019; Howanitz et al., 2019; Schmidt et al., 2021; El-Keilany et al., 2022). In the line of this research, we present the results of a project applying various CV-methods for the use case of video games, specifically Youtube video game walkthroughs for six games of the famous Soulsborne franchise developed by FromSoftware (Demon's Souls, Dark Souls 1-3, Bloodborne, Sekiro: SDT). As CV-methods, we explore movie barcodes, hue, lightness, chroma analysis and object detection via state-of-the-art algorithms and tools. Our research goals are (1) to examine if we can detect significant differences in these video games, especially across the time of release, concerning the results of CV-methods and (2) evaluate the functionality and general advantages and disadvantages of the methods for this use case.

We acquired video game walkthroughs from the platform YouTube of the channel SourceSpy91. These videos are focused on the linear playthrough of the games without commentary which results in a fitting representation of the games. The videos are in the MPEG-4 format with 1280x720 resolution, and we acquired around 105 hours of gameplay (77 GB). We manually separated the videos into the different areas/levels of the games which we regard as scenes. The number of areas ranges from 14 (Demon's Souls) to 30 (Dark Souls 2). We then generate a movie barcode for each scene via the open-source program "Movie Barcode Generator". A movie barcode is a visual representation of a video plotting the average color per time unit from left to right for one pixel. Every movie barcode has a size of 2000x720 pixels giving a representation of the color progression of the areas of the games.

We calculated hue (the general color), lightness (brightness) and chroma (saturation) values for each of the 2,000 data points per barcode, calculated the averages per game and performed statistical analysis to compare each of the games. We identified significant differences for hue ($F(5, 124) = 11.59$, $p<.001$), lightness ($F(5, 124) = 11.09$, $p<.001$) and chroma ($F(5, 124) = 9.91$, $p<.001$). Overall, the hue and lightness results point to the series becoming brighter and more colorful. These findings are also supported by qualitative-visual analysis and interpretation of the movie barcodes. The brightest game is, on average, Dark Souls 2. However, more in-depth analysis shows that this is specifically due to certain areas playing in an overly bright fantasy setting.

Object detection was performed with Detectron2, a state-of-the-art object detection model by Facebook AI that can detect up to 80 object classes like persons, animals or vehicles. Since this is a very performance intensive method, we performed object detection on one frame every 5 minutes of a game. We are currently in the process to examine the results. Via first evaluations, we identified that person detection works rather well. However, we also found multiple problems, like UI elements being classified as objects. Via domain adaption of the model with further

annotated material, we want to improve upon the basic model and compare object distributions among the games.

Overall, while our results are exploratory, we could already gain important insights via the movie barcode analysis and see a lot of potential for further research in computational game studies.

References

Arnold, T. & Tilton, L. (2019).Distant viewing: Analyzing large visual corpora. Digital Scholarship in the Humanities.

Burghardt, M., Kao, M. & Walkowski, N. O. (2018). Scalable MovieBarcodes–An Exploratory Interface for the Analysis of Movies. In IEEE VIS Workshop on Visualization for the Digital Humanities (Vol. 2).

El-Keilany, A., Schmidt, T. & Wolff, C. (2022). Distant Viewing of the Harry Potter Movies via Computer Vision. In: Proceedings of the 6th Digital Humanities in the Nordic and Baltic Countries Conference (DHNB 2022). Uppsala, Sweden.

Flueckiger, B. (2017). A Digital Humanities Approach to Film Colors. The Moving Image: The Journal of the Association of Moving Image Archivists, 17(2), 71–94. JSTOR.

Howanitz, G., Bermeitinger, B., Radisch, E., Sebastian G., Rehbein, M. & Handschuh, S. (2019). Deep Watching - Towards New Methods of Analyzing Visual Media in Cultural Studies. In Book of Abstracts of the International Digital Humanities Conference (DH 2019).

Schmidt, T., El-Keilany, A., Eger, J. & Kurek, S. (2021). Exploring Computer Vision for Film Analysis: A Case Study for Five Canonical Movies. In 2nd International Conference of the European Association for Digital Humanities (EADH 2021). Krasnoyarsk, Russia.

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[3,4], Matti La Mela[2,5], Eero Hyvönen[1,2 2]

[1]Aalto University, Finland; [2]University of Helsinki, Finland; [3]The Finnish Literature Society, Finland; [4]Finnish National Gallery, Finland; [5]Uppsala University, Sweden

9 Mar 2023
12:30-14:00

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) project[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.

Although this project is dealing with Finnish epistolary metadata (or metadata that has ended up in the Finnish archives and museums), we believe that our experiences have wider significance. In Europe, there are several digital humanities projects harvesting well-curated metadata (detailed information about senders, recipients, dates, and places) from edited letter collections - like correspSearch and Norrkor, and in some cases, the aim is to reach out to letter catalogues of CH organizations.

On the more general level, the paper participates in the ongoing discussion regarding the initial phases of data-intensive research and how this time-consuming "data work" should be described, understood, and credited. As Ahnert et al.[2] have recently argued, "the lack of discussion around such practices ... has increased mistrust of quantitative approaches in the arts and humanities. The only way to deal with this is to begin talking about the labour of cleaning and to communicate its significance as an intellectual contribution."

In the first phase of the project, we conducted a survey that was sent to over 100 CH organizations (extending from small local museums to official central archives). The paper describes how the information was collected and how the survey was constructed in order to provide us with detailed enough information regarding their 19th-century collections and metadata formats. At the same time, we had to keep the query succinct in order to make the answering as effortless as possible.

As to the data processing, we began with more than 350 000 letters, from eight 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.

A harmonizing data model for epistolary metadata collections was developed, which builds on international standards like CIDOC CRM to promote interoperability.

The most central classes are Letter, Place and Actor. Also, provenance and archival information are included.

Finally, the actor data is enriched by linking it to external databases like Wikidata and the Finnish AcademySampo and BiographySampo. 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.

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.

2 R. Ahnert, S. E. Ahnert, C. N. Coleman, S. B. Weingart, The Network Turn: Changing Perspectives in the Humanities, Cambridge University Press, 2020.

Managing Digital Humanities Data and Collections: The Records Continuum Model and the Collections of the Meertens Institute

Douwe Arjen Zeldenrust

The Royal Netherlands Academy of Arts and Sciences, The Netherlands

9 Mar 2023
14:30-16:00

Access to data and collections is one of the most fundamental starting points for every humanities researcher. Traditionally, resources are managed by archivists using the 'Life Cycle Model' (LCM). This model is viewed as fundamental to archival ideas and programs. But the growing need to work with digital records began to highlight key conceptual deficiencies in this paper-orientated model (Gilliland, 2016). Such as challenges regarding the storage of digital records across juridical and institutional borders. Consequently, the alternative 'Records Continuum Model' (RCM) was developed. The RCM is more flexible and offers insight into the complex contexts in which documents are created and managed (McKemmish, 2016). It is increasingly adopted by governments e.g. and potentially provides a viable framework for managing humanities resources as well. But within that domain, the LCM is still dominant. This paper will reflect on the potential and the issues of using the RCM as a concept for managing data and collections of institutions within the humanities.

In order to make the first steps in introducing the RCM within the domain of the humanities three case studies will be presented. These case studies originate from the collections of the Meertens Institute, a humanities institute of the Royal Netherlands Academy of Arts and Sciences. This Institute has a rich tradition in documenting and studying language and culture in the Netherlands, as well as the Dutch language and culture throughout the world. Its vast collections have been gathered over a period of over 90 years (Zeldenrust, 2020). These collections concentrate around three major topics: language variation, onomastics and ethnology. Each case study covers one of these research fields and centers around a particular collection.

The case studies have been selected on the bases of three criteria. First, the collection needs to contain both analog and digital items in order to explore the suitability of the RCM for the various characteristics of humanities collections. Second, the collection needs to have been accumulated over a long period of time in order to make the collection management policies and possible changes visible. And third, the collection must be of significant importance within the selected discipline so that it can be considered as representative of that particular field. The first case study is the audio collection 'Nederlands in Amerika' (Dutch in America) accumulated by Jo Daan (Collection 2001). The second one is the collection 'Vernoemingsnamen' (eponymized place-names) amassed by Rob Rentenaar (Collection 191). And the third case study is the 'Volkskundige Trefwoorden Catalogus' (folklore keyword list) made by Han Voskuil (Collection 141).

Lastly, the use of the RCM as a concept for managing data and collections of institutions within the humanities is part of my PhD research. As this research is ongoing, this paper will show work in progress.

References

Gilliland, Anne J. (2016). "Archival and Recordkeeping Traditions in the Multiverse and

Their Importance for Researching." In: Anne J. Gilliland, Sue McKemmish and Andrew J. Lau (eds.), Research in the Archival Multiverse. Monash University Publishing, pp 31-73.

McKemmish, Sue (2016). "Recordkeeping in the Continuum." In: Anne J Gilliland, Sue McKemmish and Andrew J Lau (eds.), Research in the Archival Multiverse. Monash University Publishing, pp 122-160.

Zeldenrust, Douwe A. (2020). Verzamelen verandert. Collectiemanagement Plan Meertens Instituut 2020 - 2023. Amsterdam: Meertens Instituut.

Archives

Meertens Institute, Royal Netherlands Academy of Arts and Sciences. Collection "Nederlands in Amerika," Collection 2001.

Meertens Institute, Royal Netherlands Academy of Arts and Sciences. Collection "Vernoemingsnamen," Collection 191.

Meertens Institute, Royal Netherlands Academy of Arts and Sciences. Collection "Volkskundige Trefwoorden Catalogus," Collection 141.

Nature and Culture in the Age of Environmental Crisis: Digital Analysis of a Global Debate in The UNESCO Courier, 1948-2011

Benjamin G. Martin[1], Fredrik Norén[2]

[1]Uppsala University, Sweden; [2]Umeå University, Sweden

8 Mar 2023
13:00-14:00

Historians have struggled to find approaches that deal adequately with the global character of twentieth-century intellectual life, in part because of the difficulty in identifying textual sources that are global in scope. One such publication is UNESCO's monthly magazine Courier. Founded in 1948 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 80s 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. The archive of this magazine was recently digitized. Working with developers at Humlab (Umeå University), we are curating this archive into a machine-readable corpus of over 1.3 million tokens, suitable for digital text analysis.

UNESCO is the United Nations organization for education, science and culture — not, in the first instance, for the environment or nature. But we have found that UNESCO's Courier reveals a striking level of interest in the natural environment, focusing on the human-nature relationship. Articles 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 particular ecosystems, and, more recently, discussed the role of human activity in changing the Earth's climate. At the heart of each of these topics was a set of fundamental questions about the relationship between nature and culture.

In this paper, we use Courier to follow a global conversation on these topics, deploying tools of digital text analysis to identify ways in which the nature-culture relationship was articulated, and to measure how the connections between concepts of nature and culture changed over time. In particular, by charting how Courier's discussion of this issue changed with the entry into UNESCO of many postcolonial states in the 1960s, we ask how understandings of the nature-culture relationship were affected by decolonization. The background to this question is the long-standing argument that non-Western cultures and indigenous peoples offer models of sustainability, in contrast to a western model of development seen as a major source of the world's environmental crisis. Courier, insofar as it promoted dialogue between the industrialized West and the "Third World" or Global South, was a key site for this global debate.

Our method in this investigation focuses on LDA topic modeling. Having computed several different models, we selected a model of 200 topics, each of which we interpreted manually and assigned 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. Topic modeling serves us, first, as a data-driven means to identify thematic foci in Courier that are not necessarily those

we were looking for. Second, we use topic networks—cooccurrences of particular topics (over a given threshold) on the same page, visualized as a network—to see how these themes interacted, and to explore the place in these networks of particular concepts. To use topic modeling for this historical investigation, we divided Courier's print-run into sub-periods and use these to measure, for example, the changing strength of natural and cultural topics in relation to historical events, such as the accession of postcolonial states to UNESCO, and milestones in the growth of public awareness of environmental issues, like the 1972 UN Conference on the Human Environment in Stockholm and the 1992 "Earth Summit" in Rio.

References

Blei, David M. "Probabilistic Topic Models." Commun. ACM 55, no. 4 (April 2012): 77–84.

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 1, 2022): 95–122.

Duedahl, Poul, ed. A History of UNESCO: Global Actions and Impacts. Basingstoke: Palgrave Macmillan, 2016.

Robin, Libby, Sverker Sörlin, and Paul Warde. The Environment: A History of the Idea. Baltimore: Johns Hopkins University Press, 2018.

Chakrabarty, Dipesh. The Climate of History in a Planetary Age. Chicago: University of Chicago Press, 2021.

Kaiser, Wolfram, and Jan-Henrik Meyer, eds. International Organizations and Environmental Protection: Conservation and Globalization in the Twentieth Century. New York: Berghahn Books, 2016.

Pernau, Margrit, and Dominic Sachsenmaier, eds. Global Conceptual History: A Reader. London: Bloomsbury, 2016.

Selcer, Perrin. The Postwar Origins of the Global Environment: How the United Nations Built Spaceship Earth. New York: Columbia University Press, 2018.

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 J. Östling, et al. Lund: Nordic Academic Press, 2020.

White, Lynn. "The Historical Roots of Our Ecologic Crisis." Science 155, no. 3767 (1967): 1203–7.

Perspectives on Sustainable Dislocated Digital Research Resources

Andrea Gasparini[1], Tom Gheldof[2]
[1]University of Oslo, Norway; [2] KU Leuven, Belgium

10 Mar 2023
14:30-16:00

As mentioned by an Oxford Digital plan (2022), significant issues in Digital Humanities have a sustainable ingredient. Firstly, technology is a problem. Software often emerges and disappears, leaving valuable data in the wild or causing costs to maintain tech competence in the organization instead of focusing on new platforms. Secondly, storing databases that have reached the end-of-life stage is still an unsolved problem. The alternative of migrating the data is not always the right solution. Thirdly, knowledge and data can be closed inside a database and often not accessible to all. Therefore, researchers should be obliged to maximize the use of their data using the FAIR principles (n.d.).

Another sustainable aspect of Digital Humanities is the impact database interfaces have in visualizing knowledge. Ensuring a rich and user-friendly interface will maximize the use. Also, when visualizing datasets, the "Interfaces become performative environments where scholars can play with the data and build their own interpretations" (Ramsay, 2011),

A suitability perspective on digital resources also includes energy consummation. This is because images and data require large amounts of power in data centers worldwide but are invisible to researchers. Another side effect is named rebound, where information of unlimited access to data storage increases the use, even the downsize for society, and the price is known (Coroama & Mattern, 2019).

The ENCODE Project

This paper presents how the large European project ENCODE (n.d.), 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 has been addressed when planning the activities and then as part of workshops and seminars that are part of the project's output. One of the outcomes of the chosen sustainability approach is a set of guidelines.

Guidelines to support sustainable dislocated digital research resources:

- Reuse of data (create a Data Management Plan)
- Public and easy access for researchers (Open Science)
- Use linked Open data
- Link to images, not copy them!
- 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)
- Decrease power usage by using links to images

- Be aware that we need to accept the scientific value of creating an open-access database!

A second output is an online course (formulated as a MOOC), developed on the platform of DariahTeach. Rather than relying on commercial technology or platforms that are maintained (and being paid for) by one of the partner's universities, we opted to choose for an already established educational platform, supported by the European consortium DARIAH.

Finally, we self-host our ENCODE database in which the innovative teaching modules, combined with the two expanded frameworks for digital competences and learning outcomes (Digcomp & Calohee), are collected.

References

Coroama, V. C., & Mattern, F. (2019). Digital Rebound—Why Digitalization Will not Redeem us our Environmental Sins. Proceedings of the 6th International Conference on ICT for Sustainability(ICT4S 2019), 2382, 31.

ENCODE. (n.d.). Bridging the <gap>in Ancient Writing Cultures: ENhance COmpetences in the Digital Era. Retrieved October 21, 2022.

FAIR. (n.d.). FAIR Principles. GO FAIR. Retrieved October 21, 2022.

Oxford University. (2022). Why Digital Sustainability Matters.

Papadopoulos, C., Rasterhoff, C., & Schreibman, S. (2022). Open Educational Resources as the Third Pillar in Project-Based Learning During COVID-19: The Case of dariahTeach. KULA knowledge creation dissemination and preservation studie, 6/1, 2–16.

Ramsay, S. (2011, January 1). On Building [Blog]. Stephen Ramsay.

Picturing Swedish Women's History: Digitizing Photographs from the KvinnSam Archives

Rachel Laura Pierce
Gothenburg University, Sweden

10 Mar 2023
14:30-16:00

Nicholson Baker's Double Fold felt like a lone digitization sceptic's jeremiad when it was published in 2002. Then, digitization was seen as a solution for the physical archive's vulnerability, a way of tearing down the institutional silos that circumscribed research, and a method of saving individual institutions loads of money. But there were socio-political and economic reasons for this initial positivism. Those working at feminist cultural heritage institutions were perhaps particularly prone to see digitization as a low-cost opportunity to integrate women and gender issues into the broad sweep of historical memory (Withers 2015, p. 136). However, as a raft of newer studies has demonstrated, the fates of the physical archive and the digital archive are interwoven (Arnold, Maples, Tilton & Wexler 2017), digital materials are interpretations of complex physical materials, not simple copies (Björk 2015), and digital documents and systems are not invulnerable to deterioration and obsolescence (Decker 2020).

There is still little research on digital collection lifecycle(s) that often include rounds of digitization, database construction and upkeep, and metadata application and restructuring. Cultural heritage institutions must understand and plan for these lifecycles if they want to construct sustainable digital and physical collections, especially because the research grants that often fund the construction of databases rarely budget for digital archive maintenance and thus do not require a full description of how this maintenance will occur (Edmond & Morselli 2020). Further, it is difficult to plan for emerging requirements. Examples abound, from institutions that must retroactively assess their digitized collections to align them with the FAIR principles to current drives to translate controlled vocabularies into linked data. What if digitization is a multiplication rather than a mitigation of the archive's/archivists' habitual problems with sustainability?

This question is particularly critical for institutions looking to ensure access to material illuminating the lives of understudied groups like women. This presentation will address the early 2000s planning for and construction of the image database at KvinnSam, the National Resource Library for Gender Research in Sweden, as well as the subsequent reassessments and potential futures of the database. KvinnSam initially digitized around 1,000 photographs, a fraction of its photographic material. The selection was driven by rights issues and high demand for images of individuals who were then attracting a lot of attention from researchers. Since then, academic research, digitization techniques, and KvinnSam itself have developed in different directions, rendering the selection, description, and database structure underlying the image database less useable. In particular, the presentation will look at how the development of KvinnSam's digital and physical archives interact over time, affecting what and how materials are accessible. How digitization is affecting the organization, description, and subsequent use of both physical and digital cultural heritage is still a question for both scholars and GLAM practitioners.

References

Arnold, Taylor, Maples. Stacey, Tilton. Lauren and Wexler, Laura. (2017). Uncovering Latent Metadata in the FSA-OWI Photographic Archive. Digital Humanities Quarterly, 11(2).

Baker, Nicholson. (2002). Double Fold: Libraries and the Assault on Paper. New York, NY: Vintage Books

Björk, Lars. (2015). How Reproductive is a Reproduction? Digital Transmission of Text-based Documents, diss. Borås: The Swedish School of Library and Information Science.

Decker, Michael J. (2020). The Finger of God is Here! Past, Present, and Future of Digital History. The Historian, 82(1), p. 7-21. DOI: 10.1080/00182370.2020.1725720.

Edmond, Jennifer & Morselli, Francesca. (2020). Sustainability of Digital Humanities Projects as a Publication and Documentation Challenge. [Sustainability of digital humanities projects] Journal of Documentation, 76(5), p. 1019-1031. DOI:

Withers, Deborah. (2015). Feminism, Digital Culture and the Politics of Transmission: Theory, Practice and Cultural Heritage. London: Rowman & Littlefield.

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
University of Regensburg, Germany

9 Mar 2023
14:30-16:00

Instagram is a growing social network [1] with a special focus on visual media. Both, internationally and in Germany politicians' and parties' political communication on Instagram has attracted researchers' interest from several fields like political science and communication science (cf.).[2]

Political communication on Instagram has been investigated differently, depending on the area of study and research interest. For the present project, we focus on content as in the campaign information and policy issues disseminated by political actors through Instagram. Our research interest was influenced by,[3] which – like most studies concerned with the political communication of election campaigns on Instagram – relies on manual analysis (e.g. through coding). In order to brace for a growing amount of visual content on social media we propose to take a computational stance: Using OCR we convert text-integrated images to computer readable text. Once the text has been digitized we are ready to apply established approaches from the computational social science (CSS),[4] like computational text analysis methods (CTAM).[5] In the scope of the present project we propose the use BERTopic[6] for topic modelling to gain insight into the policy issues covered by captions, text-integrated posts and text-integrated stories – and the share of posts or stories covering any issues at all.

We collected a sample of 2208 stories and 718 posts shared by politicians and parties in the 2021 German federal election campaign. Then we executed OCR with the help of EasyOCR and trained a Deep Learning Model to discriminate relevant from irrelevant text snippets using manually annotated data. Finally, the relevant snippets were fed into BERTopic, each post caption, text-integrated post and story as a separate document.

Once we reduced the number of topics to 15, several themes emerged. One category consisted of call-to-action content, in some cases mixed with policy issues (e.g. "Neue Wohnungen bauen [...] am 26.9. CDU wählen"), in others without (e.g. "Besser am 26.9. beide Stimmen CDU"). Another type of topic describes posts documenting the election campaign or speeches by politicians, in some cases combined with addressing the supporters and thanking them for their participation at campaign rallies. Documentation may be combined with short quotes referring to policy issues. Text-integrated images with short quotes emerged as another topic, further, we found a "change" topic, for posts and stories with parties offering to initiate abstract change when voted. Finally, in the policy-focused topics different subjects, like climate-change and economy or economy and social - blended together. We assigned a new variable to each topic classifying whether topics include mainly policy concerned content or rather policy-less documentation, call-to-actions or thanking the supporters by manually looking at a random sample for each topic.

Overall the majority of Instagram stories show, through the lens of topic modelling, a documentation of events, call-to-actions for future events and thanks to supporters.

Only a minority of stories try to disseminate policy issues. A slight majority of text-integrated posts are concerned with policy issues and captions are mostly used to promote policy issues. With triple the amount of stories to posts, the campaign appears to rather be focused on documenting events. The present analysis is only one building block towards a computational analysis of visual social media. We see future work to use speech recognition systems in order to transcribe videos posted as stories as we expect more policy issues to hide in the audio track.

References

1 Social media platforms MAU growth 2021. (n.d.). Statista. Retrieved September 1, 2022

2 Bast, J. (2021). Politicians, Parties, and Government Representatives on Instagram: A Review of Research Approaches, Usage Patterns, and Effects. Review of Communication Research, 9.

3 Haßler, J., Kümpel, A. S., & Keller, J. (2021). 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, 1–21.

4 Lazer, D., Pentland, A., Adamic, L., Aral, S., Barabasi, A.-L., Brewer, D., Christakis, N., Contractor, N., Fowler, J., Gutmann, M., Jebara, T., King, G., Macy, M., Roy, D., & Van Alstyne, M. (2009). Social science. Computational social science. Science, 323(5915), 721–723.

5 Baden, C., Pipal, C., Schoonvelde, M., & van der Velden, M. A. C. G. (2022). Three Gaps in Computational Text Analysis Methods for Social Sciences: A Research Agenda. Communication Methods and Measures, 16(1), 1–18.

6 Baden, C., Pipal, C., Schoonvelde, M., & van der Velden, M. A. C. G. (2022). Three Gaps in Computational Text Analysis Methods for Social Sciences: A Research Agenda. Communication Methods and Measures, 16(1), 1–18.

Results From Rough Data? Using Multi-Dimensional Register Analysis to Study Scottish Enlightenment Historical Writing

Aatu Liimatta, Yann Ryan, Tanja Säily, Mikko Tolonen
University of Helsinki, Finland

8 Mar 2023
13:00-14:00

While databases like Eighteenth Century Collections Online (ECCO) greatly facilitate humanities research, the low quality of the machine-readable texts produced using Optical Character Recognition (OCR) is problematic (e.g. Hill & Hengchen 2019). Bad OCR is particularly damaging for quantitative methodologies targeting the structure of the language, which require strings of consecutive words to be recognized correctly.

One such method is Multi-Dimensional Analysis (MDA), a widely-used method for register analysis, the study of language use in different situational contexts (e.g. Biber 1988). Based on the idea that linguistic features (such as grammatical constructions or word classes) are functional, MDA finds "dimensions" of co-occurring features, which are understood to co-occur because they share communicative functions. The functions of these dimensions are then analyzed, such as "informational versus involved production".

In this paper, we explore the robustness of MDA when studying functional variation in ECCO OCR data. The paper will consist of two parts. First, we will run the automated feature identification procedure for ECCO-TCP (the manually corrected subset of ECCO), and for the same texts in the ECCO-OCR. We gauge the effect of the low OCR data quality on MDA feature identification, considering the effect of the specific feature and the individual text.

We then further evaluate the applicability of MDA to ECCO-OCR data with a comparative analysis of Scottish Enlightenment authors. While some studies have applied MDA to historical texts (e.g. Biber & Finegan 1997), they have utilized smaller manually edited corpora to ensure the quality of the machine-readable texts. We will instead implement MDA on large-scale ECCO data to find out how the low quality of the data affects the results.

Our analysis examines the Enlightenment's teleological approach to 'civility': the notion, backed by writers from Adam Smith to David Hume, that societies progressed linearly from barbarity towards refinement. In the Scottish case, it was often thought of as a gradual adoption of standards of 'genteel Englishness' (Pittock: 260). These ideas pervade across a variety of texts and genres, and were key fault lines in the political differences between Tory and Whig. As the century went on, this perspective increasingly replaced the older, classical model of the world.

Our hypothesis is that this manifests in the language of the texts of the Scottish Enlightenment. We use the MDA results to compare the communicative functions of texts by Scottish authors to others. We limit the analysis to a single genre—historical writing—to control for differences in registers across genres. We look for both diachronic change and variation between 'Scottish' and 'non-Scottish' groups of authors.

We expect that many of the linguistic features commonly used in MDA will not be identified reliably in ECCO-OCR. However, we hypothesize that even a "degraded"

feature set may nonetheless provide interpretable and meaningful results, given that MDA studies have consistently produced "compatible" register dimensions regardless of the set of features analyzed (McEnery & Hardie: 115). These findings have implications for our understanding of cultural heritage collections as data, as we aim to show that even data from large-scale but error-prone databases of digitised texts have value as sources for linguistic analysis.

References

Biber, D., & Finegan, E. (1997). 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 (pp. 66-83). Société Néophilologique.

Biber, D. (1988). Variation across speech and writing. CUP.

Hill, M. J., & Hengchen, S. (2019). Quantifying the impact of dirty OCR on historical text analysis: Eighteenth Century Collections Online as a case study. Digital Scholarship in the Humanities, 34(4), 825-843.

McEnery, T., & Hardie, A. (2012). Corpus linguistics: Method, theory and practice. CUP.

Pittock, M. (2019). Historiography. In A. Broadie & C. Smith (Eds.), The Cambridge Companion to the Scottish Enlightenment (Cambridge Companions to Philosophy, pp. 248-270). CUP.

(R)Unicode: Encoding and Sustainability Issues in Runology

Elisabeth Maria Magin[1], Marcus Smith[2]

[1]University of Oslo, Norway; [2]Swedish National Heritage Board, Sweden

8 Mar 2023
14:30-16:00

This paper focuses on proposing solutions to fix the shortcomings of 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.

While most commonly associated with the Vikings in the popular consciousness, the term "runes" is used for at least four related alphabetic writing systems 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 to be found in Sweden, where the majority of them are carved into often beautifully decorated 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 Elder Futhark underwent several major changes. In the British Isles, the character inventory was expanded, while in Scandinavia, only 16 runes of the original 24 remained in use during the Viking Age. During the High Middle Ages, these 16 runes were, however, modified, and the rune row once more expanded to properly represent the phonemes in later Old West Norse.

Scholars studying runic inscriptions, therefore, by default, also study the runic writing system in question, a task which remains challenging even with digital methods to support their work. Several of the difficulties encountered can be attributed to encoding runic characters using the Unicode Runic block, which was added to the Unicode standard in 1999 to allow runes to be represented digitally. This code block leaves much to be desired for the runologist, 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.

At present, runologists 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 of the inscription. Alternatively, runologists rely on bespoke typefaces allowing them to represent a broader range of runes more accurately. 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 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 paper aims to present an overview of the current situation and suggestions for a possible solution, with particular attention paid to making use of the Unicode Form Variation Selector and Stylistic Sets in OTF fonts. Providing an introduction to

the differences between the four main runic rows on a graphemic and linguistic level, it will review the approaches to encoding runic inscriptions in relational databases and using XML, outline the issues with these current solutions and attempt to present a more sustainable approach to runic encoding better aligned with the needs of runic scholarship in the twenty-first century.

References

Fridell, Staffan, 2011, "Graphic Variation and Change in the Younger 'Futhark'" in: NOWELE 2011, Vol. 60-61; pp 69-88.

Gustavson, Helmer and Jörsäter, Steven, 1981, "Runes and the Computer" in *Michigan Germanic Studies* volume VII number 1: Proceedings of the First International Symposium on Runes and Runic Inscriptions; pp98–105. (Michigan)

Haugen, Odd Einar, 2013, "Dealing with glyphs and characters: Challenges in encoding medieval scripts" in: Document numérique 2013/3 Vol. 16, pp. 97–111. (Lavoisier, Cachan) ISSN 1279-5127 DOI 10.3166/DN.16.3.97-111

Ore, Espen S., 2002, *Runetype–Faglig Arbeidsrapport*.

Ore, Espen S. and Haavaldsen, Anne, 1997, *Runer i Bergen: Foreløpige resultater fra prosjektet "Databehandling av runeinnskrifter ved Historisk museum i Bergen". Third Edition. Website:

Ore, Espen S. and Haavaldsen, Anne 2001 "Computerising Rune-Forms" in John Higgitt, Katherine Forsyth and David N. Parsons (eds.) *Roman, Runes and Ogham: Medieval Inscriptions in the Insular World and on the Continent*. (Shaun Tyas, Donnington)

Ore, Espen S., Fiona J. Tweedie and Craig Dougan 1998 "Computers, statistics, and the grouping of rune forms" in Marilyn Deegan, Jean Anderson and Harold Short (eds.) *Digital Resources for the Humanities 1998*; pp117-127. Office for the Humanities Communications 12. (King's College, London).

Palumbo, Alessandro 2020 *Skriftsystem i förändring: En grafematisk studie av de svenska medeltida runinskrifterna*. Runrön 23. (Uppsala universitet, Uppsala) ISBN: 978-91-519-3026-8

Seim, Karin Fjellhammer 1982 *Grafematisk analyse av en del runeinnskrifter fra Bryggen i Bergen*. Dissertation. (Universitetet i Bergen)

Spurkland, Terje 1994 "K and B: One Grapheme or Two?" in James E. Kirk (ed.) *Proceedings of the 3rd International Symposium on Runes and Runic Inscriptions: Grindaheim, Norway, 8-12 August 1990*; pp269-278. *Runrön 9*. (Institutionen för nordiska språk, Uppsala Universitet, Uppsala)

Spurkland, Terje 1993 *En fonografematisk analyse av runematerialet fra Bryggen i Bergen*. PhD thesis. (Institutt for nordistikk og litteraturvitenskap, Universitetet i Oslo)

The Unicode Consortium 2003 "Archaic Scripts: Runic" in *The Unicode Standard, Version 4.0*, ch.13, 13.3, pp341–342. (Addison-Wesley, Boston) ISBN 0-321-18578-1.

The SRDM Methodology for Sustainable Semantic Infrastructure: The Case of the Wittgenstein Archives Bergen and Its Related Resources

James Matthew Fielding[1], Alois Pichler[2]

[1]Takin.solutions, Bulgaria; [2]University of Bergen, Norway

9 Mar 2023
14:30-16:00

The Semantic Reference Data Modeling method provides a means by which semantic data can be adopted by researchers in the digital humanities in an efficient way (Bruseker, Carboni Fielding, forthcoming). It aims to simplify the uptake and exploitation of semantic data by providing an intuitive approach to constructing, maintaining and sharing semantic data models. The outcome of this method - the Semantic Reference Data Models (or SRDMs) themselves - provide domain specialists with a set of mid-level models, populated with well-known entities and familiar attributes, that are of common interest across overlapping research spaces. SRDMs can thus serve as a foundation to expand beyond the target dataset and align it with related resources.

In this paper, we outline the SRDM procedure and apply it to the research assets of the Wittgenstein Archives at the University of Bergen (WAB). This process has been motivated by ongoing developments at WAB, in its mission to facilitate online access to Wittgenstein's complete Nachlass (Pichler, 2021) and its commitment to developing semantic faceted navigation of the domain. With the long-term goal of providing a framework for linking WAB data with related resources, we have chosen to adopt the CIDOC-CRM as our target ontology (https://cidoc-crm.org/). As an ISO standard for upper-level ontology in the cultural heritage field, CIDOC provides a formal structure that can serve as a common reference point among diverse actors in the field. However, as an upper-level ontology, CIDOC does not cover all cases of core concern to the archival domain in particular, let alone one that deals with specifically philosophical material (Pichler et al. 2021). Following the basic strategy of formal ontology development, we begin therefore by showing how we may map the salient elements of WAB data and its current semantic model, the "Wittgenstein ontology" (Pichler Zöllner-Weber 2013), to corresponding entity types and properties in CIDOC.

We do so by determining the ontological scope of the entities in question and formulating the typical properties employed in documenting those entities in a semantics consistent with the target ontology, and identifying those points where it becomes necessary to extend beyond the base ontological framework. The last of these, we argue, must be undertaken with care, as it potentially limits interoperability. However, following the SRDM protocol, we present a strategy for a well-documented decision chain that provides means for the identification of ontological heterogeneity and thus facilitates the measured uptake of the semantics by partner institutions. With our initial mapping of the WAB data and model in hand, we show how the structure of the SRDMs allows subsequent researchers to align adjacent information spaces modularly, by selecting only those units of documentation that they desire to use within corresponding projects, for which they can adapt ready-made, fully consistent conceptual models in a predefined formal language. Where those previously defined semantics fail to capture the particularities of the new target domain, the

process only need to be repeated, which in turn lays the foundation for the uptake of the newly defined semantic pathways and/or models by their partner institutions further down the line.

The SRDM method will not replace the need for bespoke semantic solutions, as each field of research inevitably contains idiosyncratic data that cannot be adequately covered except through the concerted efforts of those nearest to the data at hand. However, by employing the basic metadata protocol for the description of semantic patterns, which can compositionally and iteratively be applied to various data collections, the SRDM method makes semantic data modeling simpler, more consistent and above all more accessible to those who seek to benefit from reuse and redeployment.

References

Bruseker, G., Carboni, N., Fielding, J.M., Nenova, D. (forthcoming). Creating Understandable, Reusable and Sustainable Semantic Data Models: The Semantic Reference Data Modelling Method.

Pichler, A. (2021). Complementing Static Scholarly Editions with Dynamic Research Platforms: Interactive Dynamic Presentation (IDP) and Semantic Faceted Search and Browsing (SFB) for the Wittgenstein Nachlass. In: C. Navarretta and M. Eskevich (eds.), Selected Papers from the CLARIN Annual Conference 2020, pp. 194-207.

Pichler, A., Fielding, J.M., Gangopadhyay, N., Opdahl, A. (2021). Crisscross ontology: Mapping concept dynamics, competing argument and multiperspectival knowledge in philosophy. In: F. Ciracì, R. Fedriga, and C. Mararas (eds.), Quaderni de Filosophia, no. 2: Filosophia digitale, pp. 59-73.

Pichler, A., and Zöllner-Weber, A. (2013). Sharing and debating Wittgenstein by using an ontology. Literary and Linguistic Computing, vol. 28, no. 4, pp. 700-707.

Storage Over Rendition. Towards a Sustainable Infrastructure in the Digital Textual Heritage Sector

Katrine Frøkjær Baunvig, Per Møldrup-Dalum, Krista Stinne Greve Rasmussen, Kirsten Vad

Aarhus University, Denmark

9 Mar 2023
14:30-16:00

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 Baltics, as well as in the rest of Europe. One illustration: The meso-scaled project Grundtvig's Works (2009-2029) is the largest humanist project in terms of timeframe and funding on Danish ground so far. But the projects piling up on the Berlin-based Institut für Dokumentologie und Editorik's "A catalogue of Digital Scholarly Editions" 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). 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 dataset. There seems 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 – is 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 the production level (e.g., TEI; the Text Encoding Initiative) and (e.g., pre-edition compilation initiatives), the incentive structure still promote a situation of insulated digital scholarly editions focussing 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.

We hereby invite stakeholders involved in digital scholarly editing to remedy this situation. We invite stakeholders from the university and GLAM (Galleries, Libraries, Archives, Museums) sectors producing and hosting the projects, and we seek out representatives from the research foundations and the political sector financially facilitating the projects. 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 datasets, 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 for the slapdash and unambitious solutions available. Nevertheless, proper storing is the only sustainable argument for the resources going into the production of digital editions.

In sum: 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 undergird this infrastructure by making project compliance a criterium for funding.

* In a manner compliant to the principles of Findability, Accessibility, Interoperability, and Reusability (Wilkinson et al. 2016).

References

Wilkinson, M.D., M. Dumontier, I.J.Aalbersberg et al. 2016. "The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data. 3(1): 160018.

Rasmussen, K.S.G., K.S. Ravn, J. Tafdrup & K.F. Baunvig. 2022. "The Case for Scholarly Editions", DHNB 2022 Proceedings.

Stories of Sustainability

Mareike Schumacher[1], Evelyn Gius[1], Itay Marienberg-Milikowsky[2]
[1]Technical University of Darmstadt, Germany; [2]Ben-Gurion University of the Negev, Israel

8 Mar 2023
13:00-14:00

Sustainable development as a concept of fighting against climate crisis today is very present in everyday communication as well as news discourse. In addition, the term "sustainable" is used in an almost inflationary way and mostly synonymously to "long-lasting". In our contribution, we focus on the first ideas and precursors of sustainability in German and Hebrew literary texts. As the term "nachhaltend" which is German for "sustainable" was first mentioned in 1713 by Hanß Carl von Carlowitz in his Sylvicultura oeconomica Anweisung zur wilden Baum-Zucht and anchored in the very first forestry reform initiated by princess Anna Amalia of Brunswick-Wolfenbüttel in 1775 (cf. Pufé 2017, 37, Hauff 2021, 3) we take into account the closely following literary movement of German romanticism (ca. 1790-1830). In addition to the temporal proximity, this epoch is also known as being of essential relevance when it comes to descriptions of nature (cf. e.g. Bühler 2016, 85, Ulshöfer 2010, Wanning 2005, Kremer 2003, Langer et al. 2021). In our study, we found that on the one hand descriptions of nature and nature narratives are indeed very frequent in romantic texts. On the other hand ideas of sustainability are mentioned such as the principle of renewability invoked by Carlowitz (2000 [1713]). In order to find such mentions of aspects of sustainability in literary texts, we trained a classifier to automatically annotate indicators for descriptions of nature such as references to plants, animals, habitats, weather conditions etc. We then closely looked at the peaks of these indicators and identified texts in which some aspects of sustainability are interwoven into the narration. In order to find out whether German romanticism with its temporal proximity to the developments in forestry outlined above can be seen as an epoch which is especially eager to show traits of sustainability narrations, we compared a corpus of German romantic texts with a corpus of Hebrew texts which have been identified to carry some traits of romanticism too.

We thus have made a first attempt to analyse conceptions of sustainability in their historicity as a cultural phenomenon. We were able to show that ideas of sustainability show up as early as in German romanticism and thus in close temporal proximity with developments in forestry. Digital Humanities tools and methods, namely machine learning and the use of domain-adapted classifiers have proven to be of essential heuristic quality. Using these machine learning techniques we were able to take into account a comparatively large corpus of around 100 novels from romanticism and single out those of them that carry preliminary ideas on conceptions of sustainability.

References

Bühler, B. (2016) Ecocriticism eine Einführung. Edited by J.-B.-M.V. und Carl-Ernst-Poeschel-Verlag. Stuttgart: J.B. Metzler Verlag (SpringerLink: Bücher: Springer eBook Collection).

Carlowitz, H.C. von (2000) Sylvicultura oeconomica Anweisung zur wilden Baum-Zucht. Reprint der Ausg. Leipzig: Braun, 1713. TU Bergakademie (Veröffentlichungen der Bibliothek 'Georgius Agricola' der TU Bergakademie Freiberg).

Hauff, M. von (2021) Available at Nachhaltige Entwicklung Grundlagen und Umsetzung. 3., überarbeitete und erweiterte Auflage. De Gruyter Oldenbourg.

Kremer, D. (2003) Romantik. 2., überarb. und aktualisierte Aufl. Metzler (Lehrbuch Germanistik).

Langer, L. et al. (2021) 'The rise and fall of biodiversity in literature: A comprehensive quantification of historical changes in the use of vernacular labels for biological taxa in Western creative literature', People and Nature, 3(5), pp. 1093–1109.

A Sustainable West? Analyzing Clusters of Public Opinion over Time and Space in a Collection of Multilingual Newspapers (1999-2018)

Elena Fernandez Fernandez[1], Germans Savcisens[2]

[1]University of Zurich, Switzerland; [2]Technical University of Denmark, Denmark

8 Mar 2023
13:00-14:00

The United Nations 2030 Agenda for Sustainable Development, organized in seventeen sustainable development goals, is a clear indicator that signals contemporary concerns about the necessity of taking a variety of measures in the present in order to ensure a well-balanced growth of society. Discourses about sustainability, while being relatively new (Drucker), have received some critical attention across fields (Barkemeyer et al., Philippon, Lenz). However, we believe that analyzing how newspapers have encoded sustainability related concepts historically and geographically has not been yet accomplished with enough granularity.

In this paper, we will address that research gap by using a multilingual dataset of contemporary newspapers in English, German, French, Italian, and Spanish (The Times, The New York Times, Chicago Daily Herald, The Irish Times, Le Figaro, El País, NZZ, La Stampa) over an observational time of twenty years (1999-2018). Our main goal is to analyze the temporal evolution of geographic clusters of public opinion in Western societies, aiming to detect different routes of public opinion. In order to do so, we first filter our dataset using three key terms aligning with some of the seventeen United Nations Sustainable Development Goals: climate change (matching goal thirteen, "Climate Action"), pollution (matching goals six ("Clean Water and Sanitation"), and twelve ("Responsible Consumption and Production")), and environment (matching goal fifteen, "Life on Land"). We have selected these three terms as a first exploratory approach to the analysis of press coverage on sustainability discourses. Seeking to round up our investigation, we also select sustainable development, intending to inspect computationally discussions about this topic broadly speaking. To facilitate the observation of temporal variations in data behaviour, we parse our dataset in clusters of five years (1999-2003, 2004-2008, 2009-2013, 2014-2018). Then, we implement a two-step methodology consisting of Topic Modelling (Non-Negative Matrix Factorization) and Sentiment analysis, aiming to capture both the semantics and the emotions of the selected terms over time and space. Finally, we build a K-Nearest Neighbours Machine Learning Model capable of detecting proximity in both topics and sentiments and, therefore, able to provide insights about geographic and temporal variations in sustainability discourse in our selected dataset.

This paper will contribute to establish intellectual bridges between the fields of Digital Humanities and Cultural Studies and current efforts to create a sustainable world. By observing how different Eurocentric societies have decoded historically messages about sustainable development in the press, we highlight the role of both culture and zeitgeist in the depiction of a global problem that requires collective endeavours to be tackled.

References

Barkemeyer, Ralf, et al. "What the papers say: Trends in sustainability: A comparative

analysis of 115 leading national newspapers worldwide". The Journal of Corporate Citizenship, vol. 33, 2009, pp. 69–86.

Drucker, Johanna. "Sustainability and complexity: Knowledge and authority in the digital humanities". Digital Scholarship in the Humanities, vol. 36, no. 2, 2011, pp. ii86-ii94, doi: 10.1093/llc/fqab025.

Lenz, Sarah. "Is digitalization a problem solver or a fire accelerator? Situating digital technologies in sustainability discourses". Social Science Information, vol. 60, no. 2, 2021, doi: 10.1177/05390184211012179.

Philippon, Daniel J. "Sustainability and the Humanities: An Extensive Pleasure". American Literary History, vol. 24, no. 1, 2012, pp. 163-197.

Tracing the Digital Plant Humanities: Narratives of Botanical Life and Human-Flora Relationships

Paul Arthur[1], John Charles Ryan[2,3]

[1]Edith Cowan University, Australia; [2]Southern Cross University, Australia; [3]University of Notre Dame, Australia

10 Mar 2023
12:30-14:00

This paper characterises the Digital Plant Humanities (DPH) as an evolution of burgeoning environmental interest among practitioners of the Digital Humanities. DPH coalesces the theoretical and methodological frameworks of three domains: the Plant Humanities, Environmental Humanities, and Digital Humanities. After conceptualising DPH, we analyse four representative projects: (i) Native American Ethnobotany Database; (ii) Herbaria 3.0; (iii) Plant Humanities Lab; and (iv) Microcosms: A Homage to Sacred Plants of America.

In the current age of escalating environmental crisis, the future of global plant diversity is precarious. Ecologists caution that habitat degradation, land use transformations, and climatic irregularities will continue to intensify botanical extinctions. Out of this urgent context, the Plant Humanities has taken shape within the last five years as an interdisciplinary field focusing on plants and their multidimensional intersections with humankind. Entering the public domain in 2018, the term foregrounds the significance of "humanistic modes of interpretation" in the study of plants and their imbrications with people (Batsaki 2021, 2).

Plant humanists investigate the narratives and ideas linked to particular species; the designed landscapes and creative works inspired by flora; and the societal values surrounding plants. Scholarship addresses ethical concerns ranging from the social repercussions of genetically modified seeds to the moral implications of plant sentience for mainstream agriculture. Practitioners foster dialogue between anthropology, art, geography, history, literature, philosophy, plant science and other disciplines, generating novel understandings of plants and innovative models for researching them.

The Native American Ethnobotany Database (NAED) represents a predigital instantiation of the Plant Humanities. The project employs archival approaches to understand human-plant-land networks and enable cultural knowledge of plants to become more broadly available. The NAED's evolution since its analog origins in the 1970s underscores the value of long-term collaboration in the Digital Plant Humanities, exemplified, for instance, by the constellation of funding partners sustaining the project over its nearly fifty-year lifespan. Despite its durability, NAED reflects a non-participatory mode of telling plant stories. A counter-example is Herbarium 3.0, a project reflecting a user-centred ethos for disseminating plant knowledge by foregrounding plant-human interactions over time. Reconfiguring the tradition of the herbarium as a colonial institution, the project develops an interactive approach and foregrounds the significance of botanical life in an era of climatic disruption.

Released in 2021, the Plant Humanities Lab is an open-access digital platform advancing interdisciplinary approaches to plants inclusive of the arts, humanities, and sciences while showcasing the global mobilities of diverse species. Featuring interactive visualisations, the Lab sheds light on the cultural histories of plants and their role in shaping human societies. The project typifies the prevailing current emphasis

within PH on the botanical garden as "an interdisciplinary research space" co-locating humanities infrastructure, plant collections, conservation science and public engagement (Driver and Cornish 2021, sect. 3). In comparison, Microcosms bring technological innovation to the knowledge of sacred plants through the use of confocal microscopy, a specialised optical technique allowing the three-dimensional imaging of the interior of specimens. The confocal process results in vivid representations of more than seventy plants considered sacred by the Indigenous cultures of North and South America.

The paper concludes that, as a recent scholarly-activist intervention, the Digital Plant Humanities calls attention to the interconnected material, historical, cultural, and digital facets of the botanical world. DPH approaches plants as agents of nourishment, healing, and creativity vital to all aspects of human lives and livelihoods.

References

Batsaki, Yota. 2021. "Introducing the Plant Humanities Lab." Harvard Library Bulletin.

Driver, Felix, and Caroline Cornish. 2021. "Plant Humanities: Where Arts, Humanities, and Plants Meet." TEA: The Ethnobotanical Assembly 8.

GeoHumanities. 2018. "Collaborative Digital Environmental Humanities: Herbaria 3.0." GeoHumanities Forum.

Tracing the Proliferation of Socialist Realism Doctrine in Latvian Periodicals: Case Study of "Literature and Art" and "The Flag"

Anda Baklāne, Valdis Saulespurēns
National Library of Latvia, Latvia

8 Mar 2023
13:00-14:00

Goals

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 (LA, Literatūra un Māksla) and The Flag (Karogs). LA was a literary, artistic, and political weekly newspaper of the creative unions of the Latvian Socialist Republic (LSSR); The Flag was a monthly 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.

The study 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. In this study, several programmatic, ideologically saturated articles that were published in LA and The Flag in the early years were compared to the rest of the corpus to explore the proliferation and persistence of similar ideas in the course of the following decades.

Methods

In the past decade, the interest in tracking text reuse in historical newspapers has grown; notable software solutions such as Passim and BLAST have been developed. In this case study, authors have employed methodologies commonly used for plagiarism detection. As a first step, several notable articles were manually selected with the aim to compare them to all other articles in the periodicals. For each seed article, a set of candidate documents were retrieved by using n-gram comparisons. The set of candidate documents was further processed to analyse similarities; three processing methods were used: basic BoW model, vectorized TF-IDF model, and Doc2Vec embedding. Once the documents were processed, various similarity metrics were applied to determine closeness (Hamming, Canberra distance and custom distance measurements for fine-tuning). To facilitate analysis, the results were loaded into the open-source version of Neo4j graph database. Nodes were documents with corresponding metadata, and weighted edges represented different similarity measures.

Interpretation

The results of the acquired pre-processing and analysis outcomes were compared. A close reading of the set of documents that demonstrated high similarity scores to seed documents allows for further examination of the discourse, as well as critical

evaluation of the usability of similarity detection results. Although the model is based on tracking the similarity of other texts to the seed documents, the purpose of the inquiry was not to demonstrate the direct influence (or plagiarism) of these documents in particular. It is possible that individual programmatic texts played a role in the further development of the discourse, however, in this case, seed documents can be considered derivative, too, influenced by other texts that are extrinsic in regard to the corpus of LA and The Flag. Hence, the results of the analysis rather provide insight into the general prevalence of several motifs and themes.

References

Bookstein, A., Kulyukin, V.A. & Raita, T. (2002). Generalized Hamming Distance. Information Retrieval 5, 353–375.

Foltýnek, T., Meuschke, N., Gipp, B. (2019). Academic Plagiarism Detection: A Systematic Literature Review. ACM Computing Surveys, Vol. 52, Issue 6, Nov. 2020, pp. 1–42.

Gomaa W.H., Fahmy A.A. (2013). A Survey of Text Similarity Approaches. International Journal of Computer Applications (0975 –8887). Volume 68–No.13.

Pelše, S. (2003). Sociālistiskais reālisms laikrakstā "Literatūra un Māksla". Padomju perioda mākslas kritikas teorētiskie pamati. Mākslas Vēsture un Teorija [1691-0869], Nr.1, 16.-23.lpp.

Salmi, H., Paju, P., Rantala H., et.al. (2021). The reuse of texts in Finnish newspapers and journals, 1771–1920: A digital humanities perspective, Historical Methods: A Journal of Quantitative and Interdisciplinary History, 54:1, 14-28.

Zachara, M., Pałka, D. (2016). Comparison of Text-Similarity Metrics for the Purpose of Identifying Identical Web Pages During Automated Web Application Testing. In: Grzech, A., Borzemski, L., et al. (eds) Information Systems Architecture and Technology: Proceedings of 36th International Conference on Information Systems Architecture and Technology – ISAT – Part II. Advances in Intelligent Systems and Computing, vol 430. Springer, Cham.

Zeile, P. (1981). Sociālistiskais reālisms. Rīga: Liesma.

Transliteration Model for Egyptian Words

Heidi Jauhiainen, Tommi Jauhiainen
University of Helsinki, Finland

8 Mar 2023
14:30-16:00

When Egyptologists interpret hieroglyphic texts, they transliterate them with Latin letters and diacritics. A transliteration of a hieroglyphic text in, for example, a plain text file is machine-readable. However, transliteration is always an interpretation of the text, and producing it is a slow endeavor that requires checking dictionaries and sign lists. Hence, the number of openly available machine-readable hieroglyphic texts as transliteration, or in any form, is low. Computer-assisted transliteration of hieroglyphic texts will speed up producing texts for digital studies, and there have been some attempts to develop such.[1, 2] Since there still is no working automatic transliteration, this is our future aim. The transliteration method under development is based on a back-off scheme, which at its core utilizes a language model of hieroglyphic words and their transliterations together with the observed relative frequencies of the pairs. In this paper, we describe the model and how we created it using an automatic alignment method we devised based on a widely used sequence alignment algorithm.

In order to create such a model for transliteration, a corpus of machine-readable Egyptian hieroglyphic texts with their transliterations is needed. Producing hieroglyphic text with computers is not trivial, as a small sign can be placed underneath another or, for example, nested within a bigger one. Egyptologists have since the 1970s been using special text editors to encode hieroglyphic texts so that the placement of the signs is maintained.[3] The most often used encoding in such editors is the so-called Manuel de Codage (MdC). Encoded hieroglyphic texts are machine-readable, but only the pictures of the texts produced are published.[4] We have identified two sources where encoded Egyptian hieroglyphs and their transliteration pairs are available. Thesaurus Linguae Aegyptiae (TLA) includes a collection of texts where c. 280,000 Egyptian words encoded in MdC have been aligned with their transliteration counterparts.[5, 6] The second, even more extensive, source is the Ramses Transliteration Corpus (RTC) with almost 500,000 MdC encoded words.[7] The RTC consists of encoded hieroglyphic sentences, each on its own line, and respective transliteration lines in another file. However, unlike the TLA, there is no ready alignment of the MdC and its transliteration on the word level.

Original hieroglyphic texts do not include word boundaries, but since the RTC data has been made available for word searches online, it contains, in addition to texts without word boundaries, also separate versions of the files where the encoded words have been separated with underscores. 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 by the fact that many of the texts contain damaged parts. In many places, there exists a possible transliteration for these damaged parts, whether individual signs or longer passages. These guesses have been marked in a variety of ways as the transliterations have been produced by numerous scholars. Mark-Jan Nederhof previously attempted to align hieroglyphic texts using a customized scoring system to give penalties to different readings.[8] Our alignment method uses the Needleman-Wunsch sequence alignment algorithm[9] together with a

dictionary of MdC - transliteration pairs generated initially from the intact words within the TLA and completely intact lines in the RTC corpus.

After aligning the only partially intact lines of RTC, we extract the words from them and generate the MdC - Transliteration model with frequency information from all the words in both TLA and RTC. The model will be openly available as JSON files on our GitHub page. We intend to publish the scripts used in the alignment method at a later stage.

References

1 Rosmorduc, S. 2008. Automated Transliteration of Egyptian Hieroglyphs. In Strudwick, N. (ed.) Information Technology and Egyptology in 2008, Proceedings of the meeting of the Computer Working Group of the International Association of Egyptologists. Bible in Technology, 2, Gorgias Press, 167–183.

2 Rosmorduc, S. 2020. Automated Transliteration of Late Egyptian Using Neural Networks: An Experiment in "Deep Learning". Lingua Aegyptia - Journal of Egyptian Language Studies, 28, 233–257.

3 Rosmorduc, Serge. 2021. Digital Writing of Hieroglyphic Texts. In Gracia Zamacona, C. & Ortiz-García, J. (eds.), Handbook of Digital Egyptology: Texts. Monografías de Oriente Antiguo, 1. Editorial de Universidad de Alcalá.

4 Nederhof, M.-J. 2015. OCR of Handwritten Transcriptions of Ancient Egyptian Hieroglyphic Text. In Proceedings of Altertumswissenschaften in a Digital Age: Egyptology, Papyrology and Beyond.

5 Teilauszug der Datenbank des Vorhabens "Strukturen und Transformationen des Wortschatzes der ägyptischen Sprache" vom Januar 2018. Akademienvorhaben Strukturen und Transformationen des Wortschatzes der ägyptischen Sprache. Text-und Wissenskultur im alten Ägypten. 2018. urn:nbn:de:kobv:b4-opus4-29190.

6 Schweitzer, S. 2021. AES - Ancient Egyptian Sentences; Corpus of Ancient Egyptian sentences for corpus-linguistic research. GitHub.

7 Rosmorduc, S. 2021. Ramses automated translitteration software. In Lingua Aegyptia (2021-06-15, Vol. 28, pp. 233–257). Zenodo.

8 Nederhof, M.-J. 2009, Automatic alignment of hieroglyphs and transliteration. In Strudwick, N. (ed.), Information Technology and Egyptology in 2008: Proceedings of the meeting of the Computer Working Group of the International Association of Egyptologists. Bible in Technology, 2, Gorgias Press, 71-92.

9 Needleman, S. B. and Wunsch, C. D. (1970). A General Method Applicable to the Search for Similarities in the Amino Acid Sequence of Two Proteins. Journal of Molecular Biology, 48(3), 443–453.

Understanding Without Knowing: Livonian Intangible Cultural Heritage

Valts Ernštreits
University of Latvia, Latvia

10 Mar 2023
12:30-14:00

Intangible cultural heritage (ICH) documented in a particular language may only be accessed by those proficient in that language or by using human or machine translation support. As endangered languages tend to have a limited number of individuals proficient in these languages and proficiency within the research community is mainly possessed by linguists due to their own personal and research interest, the non-proficient parts of the community as well as researchers from other fields, are excluded from having access to this ICH. The current article focuses on approaches to overcome this language barrier to provide access to the ICH of Latvia's indigenous Livonian community.

The majority of Livonian ICH is collected in Livonian, which is one of the world's most endangered languages, spoken by a little over 30 people, which includes linguists. Since 2017, work has started on the creation of digital linguistic resources for Livonian, primarily from a linguistic perspective and initially based on the Livonian-Estonian-Latvian dictionary (LELS). The current cluster of resources includes a collection of Livonian vocabulary and place names, a collection of morphological data, and morphologically annotated corpora, which were initially intended to be a source of lexical and morphological forms. As more texts (folk tales, books, periodicals, etc.) have been added to the corpus, it becomes evident that along with its linguistic value, it is developing into a collection of Livonian ICH. However, being collected in Livonian, this ICH is not accessible to the majority of the Livonian community, researchers, and others who are not proficient in Livonian. This creates a demand for a translation: the Livonian community would need a translation into Latvian, while researchers, as well as others interested in this material, would need one into English.

In a perfect scenario, all texts would be translated by human professionals; however, this is not a viable option due to the complexity of the task, limited resources (both fiscal and human), as well as the length of time involved. Possibilities for providing access to the ICH using automated approaches are currently being explored, including the possibility of automated solutions based on existing and new multilingual data. Two types of solutions are being developed.

At the first stage – providing a basic idea about the content of the texts (e.g., for selecting the necessary examples) – technically simple solutions based on ones that already exist are being explored. Even basic multilingual information may be sufficient for this task (e.g., providing the ability to search for keywords in another language). This can be accomplished by using automated identification of lemmas using corpus annotations, statistics, and morphological analysis and subsequently providing possible translations of these lemmas from multilingual dictionaries.

At the second stage – translation of the content – the use of NLP-based machine translation is being explored (see Ernštreits, Fišel et al 2022). NLP modules are trained with parallel corpora formed by the multilingual (Livonian, Estonian, Latvian; English being added) dictionary and its examples, other parallel and monolingual

data in Livonian and its related and contact languages (Estonian, Finnish, Latvian; see Rikters et al ACL 2022). The main limitation is, however, posed by a lack of sufficient multilingual data, but this issue can be partially resolved by using lemma-based multilingual data produced during the first stage.

An ideal automatic translation of Livonian is not likely to be achieved anytime soon, if ever, but if these solutions work and can be successfully applied, they might provide new opportunities both to the ICH community and, most definitely, to researchers.

References

Valts Ernštreits, Mark Fišel, Matīss Rikters, Marili Tomingas and Tuuli Tuisk. 2022. Language resources and tools for Livonian. Eesti Ja Soome-Ugri Keeleteaduse Ajakiri. Journal of Estonian and Finno-Ugric Linguistics, 13(1), 13–36.

Matīss Rikters, Marili Tomingas, Tuuli Tuisk, Valts Ernštreits, and Mark Fišel. 2022. Machine Translation for Livonian: Catering to 20 Speakers. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pages 508–514, Dublin, Ireland. Association for Computational Linguistics.

LELS = Tiit-Rein Viitso and Valts Ernštreits. 2012. Līvõkīel-ēstikīel-leţkīel sõnārõntõz. Liivi-eesti-läti sõnaraamat. Lībiešu-igauņu-latviešu vārdnīca. Tartu, Rīga: Tartu Ülikool, Latviešu valodas aģentūra.

Untapped Data Resources: Applying NER for Historical Archival Records of State Authorities

Venla Räsänen, Tanja Välisalo, Ida Toivanen, Jari Lindroos, Antero Holmila, Jari Ojala

University of Jyvaskyla, Finland

10 Mar 2023
12:30-14:00

Archives around the world are digitising their material at a vastly growing speed. National archives are no exception to this task and have been digitising various archival materials. While digitisation has become a more frequent practice, the quality of the digitisation varies immensely between archives and different times of digitisation. This presents a set of problems for the use of the digitised data. Many archives have solved this problem by saving the physical archives as well.

The National Archives of Finland started mass digitisation in 2019. Mass digitisation aims to digitise vast amounts of materials from state authorities so that the original documents can be destroyed after digitisation (kansallisarkisto.fi). This means that massive amounts of records will be made available to the researchers in various fields of study. This opens up a wide range of possibilities for researchers. For historical research in particular, 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 (Fridlund, Oiva & Paju 2018).

The danger of massdigitatisation lies in the user experience: creating large databases which are difficult to access and use is a waste of resources. Making digitised archives useful and usable for research purposes demands enriching the data in various ways. One way to make the data more approachable for researchers is to use a natural language processing (NLP) task called Named Entity Recognition (NER).

In this paper, we will consider what are the potential benefits of using named entity recognition in historical research with digitised state authority archives. Named entity recognition (NER) was originally developed as a form of information extraction (IE) (Palmer & Day 1997). The core task of NER is locating and naming predefined entities (Humbel et al. 2021). Compared to common search results, using NER based search reveals a wider set of results. Additionally NER can be used as a methodological tool for history research. Whether it is mapping persons, locations or other wanted focus entities, NER can help the researcher do a variety of deductions based on the results (Fields et al. 2022).

In practice, using the digitised material from the Ministry of Economic Affairs and Employment, we will explore the utilisation of NER with the archival records of the state authorities. To deepen our understanding of potential needs for NER based tool development, we are conducting an online survey targeted at researchers interested in using state authority archives. The survey is distributed to researchers in the fields of history and social sciences in particular, through universities and research conferences in Finland during autumn 2022. The survey results will be applied to support NER development in a sustainable way, designing it from the start to be useful for diverse research perspectives.

This work is a part of the FIN-CLARIAH infrastructure program 2022-2023, which aims to develop processes and methods for processing unstructured text in social sciences and humanities.

References

Humbel, Marco, Nyhan, Julianne & Vlachidis, Andreas (2021). "Named-entity recognition for early modern textual documents: a review of capabilities and challenges with strategies for the future." Journal of Documentation 77, 6.

Palmer, David D. & Day, David S. (1997) "A Statistical Profile of the Named Entity Task. In Proceedings of the ACL Conference for Applied Natural Language Processing."

Petri Paju, Mila Oiva, & Mats Fridlund (2020). Digital Histories: Emergent Approaches Within the New Digital History. Helsinki University Press.

Fields, Sam, Cole, Camille, Oei, Catherine & Chen, Annie. "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.

Vectors of Violence: Changing Conceptualizations of Terror, Terrorism and Våldsbejakande Extremism in Swedish Parliamentary Data, 1971–2018

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

University of Gothenburg, Sweden

9 Mar 2023
12:30-14:00

This paper examines conceptual changes in the discourse on terrorism and political violence in the context of Swedish parliamentary deliberations 1971–2018. Analysing debate transcripts (protokoll) from the Swedish Parliament, the study forms part of the major mixed methods project SweTerror (2021–2024, see Edlund et al. 2022), that draws on an array of theoretical and methodological considerations from, among others, linguistics, terrorism studies, political science and language technology.

The focus of this study is examining the altering ways of conceptualising political violence. In the time period preceding 1971, political violence was only occasionally verbalized as terrorism, more frequently as terror as well as by means of other lexical resources. The highly contested term terrorism, in turn, was more often used for signifying acts committed by states in the sense of "state terrorism". However, around 1970, a terminological shift can be observed in so far as terrorism increasingly became deployed with reference to political violence committed by clandestine groups, thereby perpetrating acts of symbolic value (cf. Jackson 2011). Thus, the question arises to which extent this lexical shift is indicative of discursive changes concerning the topic of political violence.

This study has a twofold aim. First, we will analyse the semantic shift from terror to terrorism by deploying language technology techniques, including word vectors which can give the contextual closeness of words (semantically similar words having similar vectors). This allows us to capture any change of usage (i.e. comparing the words' word vectors over time when new terms are introduced). By comparing similar contexts, we can predict the words with similar meanings but also find what words are used in the same contexts earlier or later on. Combining this with other aggregation and partitioning techniques, we get, by filtering on these features and the text metadata, e.g. gender, seniority or party affiliation, a powerful drill-down and navigation feature for exploration.

Second, we will analyse another significant shift in the Swedish discourse history of terrorism, occurring 2014–2015 when the term våldsbejakande extremism ('violent extremism') was introduced in the context of violent Jihadism and radicalization. Andersson (2018) argues that våldsbejakande extremism has to a large extent, replaced terrorism in the political-juridical discourse on politically and religiously motivated violence. However, it yet remains to be explored whether this observation can be quantitatively substantiated on the basis of larger datasets. We, therefore, examine the introduction and distribution of våldsbejakande extremism in the Parliamentary debate transcripts and its word vectors in comparison to those of terrorism in the same period of time.

Consequently, the paper elaborates on the extent to which terror, terrorism and våldsbejakande extremism have been used to denote different, similar or related activities and/or stances. We also map what fields of action or political, ideological

domains that these concepts evoke and to what strands of discourse on political violence they relate and contribute to.

References

Andersson, Dan-Henrik (2018): "Från terrorism till våldsbejakande extremism". In: Malin Arvidsson/Lena Halldenius/Lina Sturfelt (eds.): Mänskliga rättigheter i samhället. Malmö: Bokbox, 149–164.

Edlund, Jens et al. (2022): "A Multimodal Digital Humanities Study of Terrorism in Swedish Politics". In: Kohei Arai (ed.): Intelligent Systems and Applications. Proceedings of the 2021 Intelligent Systems Conference (IntelliSys) Volume 2. Springer, 435–449.

Jackson, Richard (2011): "In defence of 'terrorism'". Behavioral Sciences of Terrorism and Political Aggression 3 (2), 116–130.

Magnus P. Ängsal, Daniel Brodén, Mats Fridlund, Leif-Jöran Olsson, & Patrik Öhberg, "Linguistic Framing of Political Terror: Distant and Close Readings of the Discourse on Terrorism in the Swedish Parliament 1993–2018", in Tomaž Erjavec & Maria Eskevich, eds. CLARIN Annual Conference Proceedings 2022, Series CLARIN Annual Conference Proceedings (Prague: CLARIN, 2022), 69–72.

Patrik Öhberg, Daniel Brodén, Mats Fridlund, Victor Wåhlstrand Skärström & Magnus P. Ängsal, "Unifying or Divisive Threats? Anxiety about Political Terrorism and Extremism among the Swedish Public and Parliamentarians, 1986–2020", in Karl Berglund, Matti La Mela & Inge Zwart, eds., DHNB 2022: Proceedings of the 6th Digital Humanities in the Nordic and Baltic Countries Conference (DHNB 2022), Uppsala, Sweden, March 15-18, 2022, CEUR-WS vol. 3232 (Aachen: CEUR-WS.org, 2022), 145–158.

Mats Fridlund, Daniel Brodén, Leif-Jöran Olsson & Magnus Ängsal, "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 Digital Parliamentary Data in Action (DiPaDA 2022) Workshop Co-located with the 6th Digital Humanities in the Nordic and Baltic Countries Conference (DHNB 2022), Uppsala, Sweden, March 15, 2022, CEUR-WS vol. 3133 (Aachen: CEUR-WS.org, 2022), 167–175.

Jens Edlund, Daniel Brodén, Mats Fridlund, Cecilia Lindhé, Leif-Jöran Olsson, Magnus Ängsal & Patrik Öhberg, "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: Proceedings of the Intelligent Systems Conference (IntelliSys) 2021, Vol. 2, Lecture Notes in Networks and Systems 295 (Cham: Springer, 2022), 435–449.

Våran Klubb, Barnvaktsklubben or The Baby-Sitters Club: The Data-Sitters Club as an International Community

Agnieszka Backman[1], Quinn Dombrowski[2]

[1]Uppsala University, Sweden; [2]Stanford University, United States of America

10 Mar 2023
14:30-16:00

Founded in late 2019, the Data-Sitters Club is a feminist digital humanities collective that takes Ann M. Martin's 1980s and 1990s girls' book series The Baby-Sitters Club as a corpus for exploring computational text analysis methods, writing up the code, results, and the entire decision-making process using colloquial language. The goal of the Data-Sitters Club is to create an accessible resource where anyone can learn text analysis methods, and understand the debates, frustrations, and compromises that come with doing this work. Over the past three years, the group's six founding members have worked with "guest data-sitters" who are specialists in particular methods to write 17 "books" on topics including creating a corpus, text comparison algorithms, machine learning, TEI, AntConc, sentiment analysis, principal component analysis, and text classification.

At the same time, the group is keenly aware of the fact that pedagogical resources for "text analysis" often encompass only "English text analysis". Languages with non-trivial inflection typically need to be pre-processed before one can attempt "simple" word count tools like AntConc or Voyant, or any method that involves word frequencies (Dombrowski 2020). Sentiment analysis requires language-specific vocabularies or language models, as do syntactical analyses. The performance of widely-used language models for English tends to be far better than the equivalent models for other languages. Most general-purpose guides to "text analysis" elide these language-specific issues, gesturing to them in passing, if at all. For this reason, the Data-Sitters Club started a "Multilingual Mystery" sub-series to explore The Baby-Sitters Club series in translation.

In this talk, we will discuss the Data-Sitters Club as an international community that spans institutional affiliations, roles (e.g. our group includes faculty, staff, students, postdocs, and at times even children), countries, and languages. Our data set – in-copyright books from many publishers, some no longer extant – is challenging to work with, not least because we cannot make it open, and we face barriers to access in different countries. This talk will touch on the complexity of copyright and data law when working internationally, specifically the delays in obtaining a Swedish corpus due to uncertainty around the legal situation in Sweden before the implementation of the European Union Data and Text Mining Directive on August 1, 2022. Finally, we will focus on a case study of a Data-Sitters Club Multilingual Mystery exploring translation strategies in the Swedish corpus, using the Bleualign algorithm (Senrich 2011) for text alignment with English, along with the Swedish spacCy NLP model. Since the publishing history of the 17 translated books spans two different publishers, with five translators working in three distinct phases over ten years, this facilitates investigation into which norms originate with the publisher and which come from the different translators. The first book has been published twice, using the same translation but with changes to it the second time, and can thus serve as a model for the second publisher's guidelines. Personal names, place names, and titles will be the focus of this case study, considering how they reflect either a source or target language-focused translation.

We hope that the Data-Sitters Club can serve as a model for future projects that draw on popular culture (and its translation) as an entry point for making computational methods more engaging and accessible for people with a broader range of backgrounds and interests.

References

Dombrowski 2020. Preparing Non-English Texts for Computational Analysis Modern Languages Open, (1), 45.

Sennrich & Volk 2011. Iterative, MT-based sentence alignment of parallel texts. In: NODALIDA 2011, Nordic Conference of Computational Linguistics, Riga.

Wikidata for Authority Control: Sharing Museum Knowledge With the World

Alicia Fagerving
Wikimedia Sverige, Sweden

9 Mar 2023
12:30-14:00

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. In this presentation, we will introduce the project Wikidata for Authority Control, in which the Nationalmuseum and the National Historical Museums, in partnership with Wikimedia Sverige, a regional chapter of the Wikimedia Foundation, explore this new area.

The aim of our project is to develop and evaluate methods of linking museums' 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.

The participants will see an example of how a major national cultural heritage organization can carry out a Wikidata-focused project. They will gain an insight into what is required to start working with Wikidata and understand its community, and what tools are useful. They will also learn how a cultural heritage organization can benefit from linking their data to Wikidata, providing arguments to use when initiating similar partnerships of their own.

Our project is spanning three years, and apart from authority files, work with geographical thesauri, iconography and other vocabularies has been done. A desired outcome is not only that more of the museums' open data is available via Wikidata, but also that the project members have developed relevant skills to work with Wikidata and OpenRefine. In-house Wikidata expertise benefits the institution and can be shared with other actors.

Show-and-Tells

Air Pollution Affecting the Speech Complexity of Finnish Members of Parliament

Anna Kristiina Ristilä
University of Turku, Finland

This study was a simplified replication of a study made in Canada, where air pollution was compared to politicians' speech complexity (Heyes et al. 2019). The pollution particles measured were of size PM2.5, extremely fine particles which cannot normally be seen, can permeate most commercial air filters and get deep into the lungs and even into the bloodstream. The data used in this study was Finnish parliamentary plenary speeches from the years 2006-2011 as well as daily PM2.5 concentration data from the same years measured near the Parliament House in central Helsinki.

Speech complexity was represented by Flesch-Kincaid grade level indexes, which were calculated similarly as in the original study. This index uses the number of sentences, syllables and words in a speech, so automatic syllabification was conducted with the Python FinnSyll module. There was no PM2.5 data for some days, so speeches from those dates were also dropped.

The original study by Heyes et al. showed how air pollution caused politicians to speak less, in shorter sentences and with less complex words during high pollution. Similar results were gained from this study.

The comparison of the Finnish results to the results derived from Canada is of interest since both Ottawa and Helsinki are relatively low-pollution cities overall. The similar results strengthen the understanding that even in relatively clean cities (globally speaking) air pollution can cause measurable negative cognitive and health impairment.

References

Archsmith, James, Anthony Heyes, and Soodeh Saberian. 2018. "Air Quality and Error Quantity: Pollution and Performance in a High-Skilled, Quality-Focused Occupation." Journal of the Association of Environmental and Resource Economists 5 (4): 827–63.

Bellani, Luna, Stefano Ceolotto, Benjamin Elsner, and Nico Pestel. 2021. "Air Pollution Affects Decision-Making: Evidence from the Ballot Box." SSRN Electronic Journal.

Bondy, Malvina. 2020. "Crime Is in the Air: The Contemporaneous Relationship between Air Pollution and Crime." Journal of the Association of Environmental and Resource Economists 7 (3): 555–85.

Dechezleprêtre, Antoine, Nicholas Rivers, and Balazs Stadler. 2019. "The Economic Cost of Air Pollution: Evidence from Europe." OECD Economics Department Working Papers 1584. Vol. 1584. OECD Economics Department Working Papers.

Heyes, Anthony, Nicholas Rivers, and Brandon Schaufele. 2019. "Pollution and Politician Productivity: The Effect of PM on MPs." Land Economics 95 (2): 157–73.

"How Does PM Affect Human Health?" n.d. United States Environmental Protection Agency. Accessed January 5, 2023.

Lehtomäki, Heli, Camilla Geels, Jørgen Brandt, Shilpa Rao, Katarina Yaramenka, Stefan Åström, Mikael Skou Andersen, Lise M. Frohn, Ulas Im, and Otto Hänninen. 2020. "Deaths Attributable to Air Pollution in Nordic Countries: Disparities in the Estimates." Atmosphere 11 (5): 467.

Lehtomäki, Heli, Antti Korhonen, Arja Asikainen, Niko Karvosenoja, Kaarle Kupiainen, Ville-Veikko Paunu, Mikko Savolahti, et al. 2018. “Health Impacts of Ambient Air Pollution in Finland.”nternational Journal of Environmental Research and Public Health 15 (4): 736.

Lu, Jackson G. 2020. “Air Pollution: A Systematic Review of Its Psychological, Economic, and Social Effects.” Current Opinion in Psychology 32 (April): 52–65.

Collecting and Sharing References in Christian-Muslim Religious Encounters With the OTRA Framework: An Assessment and a Roadmap

Jacob Langeloh
University of Copenhagen, Denmark

Unsurprisingly, texts in the tradition of Christian-Muslim religious encounters (CMRE) are full of cross-references. To discuss the 'other', one referred to essential sources, such as the Bible and the Qur'an, and to their interpretation, and thus a thick network of cross-references emerges. To discover these connections, however, a researcher has to be lucky – perhaps someone else left a hint – or sift through large parts of the tradition. With the OTRA framework (Ontology for the Transmission and Re-Use of Argumentative Patterns), I propose a different approach that builds on the communal engagement and sustainable sharing of research results. In this show and tell, I introduce the general idea of the OTRA framework and outline the measures undertaken to make this a communal effort within the CMRE research community.

OTRA builds on the observation that the relationship between references and argumentation is fluid. The same reference can support different conclusions, and the same conclusion can build on different references. The ontology, therefore, contains categories to describe references, re-uses, and argumentations, that can then be analyzed to describe changes in argumentation over time. OTRA is based on CIDOC CRM's LRMoo extension[1] and it incorporates and homogenizes parts of the Sharing Ancient Wisdoms (SAWS)[2] and the Hypermedia Dante Network (HDN)[3] vocabularies.

In the first part of my contribution, I provide a quick overview of this approach. In the second part, I will address questions about communal data curation. To enter a significant part of the field that allows sensible conclusions, contributions from specialist researchers, who are not DH oriented, are needed. I present various measures undertaken in order to support such participation – such as providing an easy interface or rewards for entering data – and question their effectiveness.

References

1 International Working Group on LRM, FRBR and CIDOC CRM Harmonisation: https://cidoc-crm.org/ModelVersion/lrmoo-f.k.a.-frbroo-v.0.7 [accessed 3.1.2022].

2 Valentina Bartalesi, Nicolò Pratelli, Carlo Meghini, Daniele Metilli, Gaia Tomazzoli, Leyla M G Livraghi, Michelangelo Zaccarello, A formal representation of the divine comedy's primary sources: The Hypermedia Dante Network ontology, Digital Scholarship in the Humanities, Volume 37, Issue 3, September 2022, Pages 630–643.

3 Mark Hedges; Anna Jordanous; K. Faith Lawrence; Charlotte Roueché; Charlotte Tupman - Computer - Assisted Processing of Intertextuality in Ancient Languages, jdmdh:1375 - Journal of Data Mining Digital Humanities, 7 septembre 2017, Numéro spécial sur le traitement assisté par ordinateur de l'intertextualité dans les langues anciennes.

Creating a Crowdsourcing System and Community to Collect Information on Local Everyday Objects

Sakiko Kawabe

National Museum of Japanese History, Japan

This presentation will introduce a newly-started project for developing a crowdsourcing system to help local museums collect information on their collections of everyday objects in Japan. Everyday objects have been collected and preserved as historical and cultural materials of the regions. However, many local museums are poor in detailed records of these objects in their catalogs; for example, some objects lack essential information on how and for what they were used or who made, used, and donated them and when. Such undescribed collections cannot be utilized in museum exhibitions or other educational or scholarly activities but just be abandoned.

To solve this problem, the Academic Repository Network (Re*poN) started a project to create a crowdsourcing system to gather information on everyday objects in cooperation with Shibetsu City Museum in Hokkaido. In early October 2022, we held a first hybrid meeting in the city, aiming to know how effectively we can collect information on these objects when informants gather from inside and outside the region in person and virtually. The offline participants were local museum staff, researchers (from the fields of informatics, museology, or ethnography), and local senior citizens who may know about these old everyday objects, while museum professionals and researchers from outside the city also joined online. In the meeting, we take a look at undescribed everyday objects from the museum collections one by one. At the same time, participants discussed and recorded on google forms what these objects are and when, where, and how these objects were made and used, sharing any related information from their own experience, the internet, or any other sources.

Through the meeting, we learned the importance of combining online and offline and the effectiveness of creating a community beyond the region to collect information on everyday objects.

References

Kawabe, S. 2022. Loose Preservation of Mingu: from a Perspective of Lifecycle of Things. Journal of Rural Planning Association. 40(1). 6-9.

Kawabe, S. 2021. Everyday Object Collections Formed by Collectors and Contributors in the Local Living Context: An investigation on background of the collecting and functions of collected objects in the Noto Peninsula, Japan and Ifugao Province, Philippines. Graduate School of Kanazawa University (doctoral thesis).

Digitisation of Printed Books in Old Latvian Orthography for the Preservation and Sustainability of Cultural Heritage: Workflow and Methodology

Karina Šķirmante, Silga Sviķe
Ventspils University of Applied Sciences, Latvia

To study the first names of organisms and other botany, zoology, and mineralogy terms introduced into the Latvian and their changes through time, the old printed books should still be used since not all data are machine-readable. Therefore, in order to ensure the sustainability of such research, this study describes the digitisation of the book of G. H. Kawall's Dieva radījumi pasaulē (God's Creatures in the World), 1860, translated from German into Latvian, in which a lot of the very first natural science terms (e.g. augs (plant), augu valsts (plant kingdom)) appeared in Latvian.

Kawall's work is an important cultural heritage for studies of terminology in the natural sciences. The digitised text allows extracting the terms and add them to a collection of names of organisms newly developed as an interactive data collection and management system for research purposes, especially terminology change studies, including provisions of sustainable and modern processing of such data.

The goal of this study is to provide insight into the workflow and methodology for the recognition and digitisation of old Latvian orthography by using algorithms of optical character recognition (OCR) based on the machine learning TesseractOCR engine. It was necessary to train the machine learning model to recognise individual features of the old Latvian orthography. The alphabet of the old Latvian orthography available on the website of the National Library of Latvia was used. A data set of the book of Aronu Matīss (1858–1939) Aronu Matisa Vecais Pantenius was used as the first training data. The developed solution for the first data set achieved 82% effectiveness and helped in making the data processing more efficient. In the report, the authors will discuss the problems in the digitisation and processing of text fragments from the main data set and will present the developed solution.

References

Jasmonts, G., Sviķe, S. Šķirmante, K. (2022). New Information Extracting and Analysis Methodology for the Terminology Research Purposes: The Field of Biology. CEUR Workshop Proceedings Volume 3160. IRCDL 2022: 18th Italian Research Conference on Digital Libraries.

Dos and Don'ts of Building a Pan-European Biographical Knowledge Graph: Statistical Analysis of the InTaVia-Platform

Matthias Schlögl[1], Joonas Kesäniemi[2], Jouni Tuominen[2, 3], Victor de Boer[4], Go Sugimoto[4], Carla Ebel[1]

[1]Austrian Academy of Sciences, Austria; [2]Aalto University, Finland; [3]University of Helsinki, Finland; [4]Vrije Universiteit Amsterdam, The Netherlands

"In/Tangible European Heritage – Visual Analysis, Curation and Communication" (InTaVia) aims at bringing together data from biographical dictionaries and cultural heritage. The basic layer of the InTaVia Knowledge Graph (IKG) is built using structured data from 4 European National Biographies – Austria, Finland, The Netherlands and Slovenia. These datasources have been converted using the IDM-RDF (https://github.com/InTaVia/idm-rdf) and integrated in one knowledge graph. IDM-RDF uses CIDOC CRM as a basis, the Bio CRM extension (especially for person-specific attributes), the proxy solution developed in Open Archives Initiatives ORE and PROV-O for data provenance.

On top of the first layer a second level layers are built enriching the base layer with data from reference resources such as Wikidata. Currently second level layers exist for places and cultural heritage objects (CHO) from Wikidata. CHO data from Europeana is being worked on.

Currently the IKG contains data on around 300.000 distinct persons from various periods that had an impact on the history of European countries in one way or the other. Additionally it contains data on places, institutions, events and – pulled from Wikidata – cultural heritage objects these persons interacted with. However, the IKG has to tackle several problems: the source data is quite unbalanced – while some persons have connections to 50 and even more events, others don't even have birth or death events attached, the sources contain biased data in various forms – women are constantly underrepresented and the biographies have historically been used to form "national myths" – and other problems.

To tackle these problems and provide researchers with the additional information on the data in use a quantitative statistical analysis has been conducted. This contribution presents the results of the analyses, discusses possible solutions to the revealed shortcomings and shares the "lessons learned" when integrating large historical secondary data sources.

Environmental Concerns in COVID-19 Vaccine Discussions on Twitter: Between Science Enthusiasm and Science Denial

Jana Sverdljuk[1], Bastiaan Bruinsma[2]

[1]University of Agder, Norway; [2]Chalmers University of Technology, Sweden

Over the last few years, the COVID-19 pandemic dominated the public debate. With this came a decrease in attention to other issues, such as environmental protection. Yet, recently, critical studies linking the pandemic with such issues have begun to emerge. These often point to the impact of the production and distribution of vaccines on not only the local but also the global environment. For example, Hasija et al. (2021) point out that the disposal of materials needed for vaccination often takes place at sea. While the precise effects of this are unknown, the effects of used (bio)medical materials can be damaging.

Here, we aim to better understand the global public media discourse surrounding the environmental impact of vaccines. This, as this debate combines various aspects of a wider cultural war, in which rationality and trust in science are under suspicion. Both vaccines and the environment were often at the centre of these debates, and as such, the study of how these two were intertwined during the pandemic can lead to new insights into whether positions in this debate lean towards a pro-science or anti-science stance.

To study the debate on the vaccines and environment, we make use of a corpus containing more than 68 million Tweets mentioning "vaccine" between 2018 and 2022. The size and scope of this data set not only allow us to get a wide perspective of the debate but also to see how the pandemic altered it. For example, when we find a mention of "environment", are these the voices of the science proponents and environmentalists, or those of vaccine sceptics and science deniers? Also, to what degree is the possible negative impact of vaccine on the environment used by the latter as an argument against vaccines?

To answer these questions, we will combine a Structural Topic Model (STM) (Blei 2012) with Critical Discourse Analysis (CDA) (Fairclough 1995; Wodak 2001). This allows us to generate an initial structure in our data using STM, which CDA can then use to identify discourses and subject positions. Further using STM, we can then "track" these positions over time and between users and see how external events shaped them. In all, it allows us to present a more complete picture of the interrelationship between vaccines and the environment and how they exist in the public debate.

References

Blei, D. M. (2012). "Probabilistic Topic Models". Communications of The ACM, 55 (4): 77–84. doi: 10.1145/2133806.2133826.

Fairclough, N. (1995) Critical Discourse Analysis: The Critical Study of Language. Longman

Hasija V., Patial S., Raizada P., Thakur S., Singh P., Hussain C.M. "The environmental impact of mass coronavirus vaccinations: A point of view on huge COVID-19 vaccine waste across the globe during ongoing vaccine campaigns." Sci Total Environ. 2022 Mar 20;813:151881. doi: 10.1016/j.scitotenv.2021.151881.

Wodak, R, Michael Meyer (2001). Methods of Critical Discourse Analysis. Sage Publications

Making Modern Drama: Change Detection in the Information Space of Henrik Ibsen's Writing

Ellen Rees[1], Frida Hæstrup[2], Kristoffer Laigaard Nielbo[2]
[1]University of Oslo, Norway; [2]Aarhus University, Denmark

A growing scientific literature attempts to describe the change in cultural and historical information systems based on the relationship between new information and information longevity.[1, 2, 3]. These approaches are fundamentally information-theoretical and formally model complex cultural dynamics with windowed relative entropy in order to describe the degree of "surprise" in probabilistic representations of cultural objects.[1] "Surprise" is, in this context, modeled both in relation to the past and the future. "Novelty" is a measure of how surprising the information patterns in, for instance, a document, are, given past documents. "Transience" a measure of the extent to which those patterns persist in future documents, and "resonance" a measure of the extent to which patterns in future documents conform to the document novelty. One study has shown that a similar information-theoretical approach is highly productive for profiling the Danish pastor, author, and poet N.F.S. Grundtvig (1783-1872).[4]

In this study, we encode Norwegian playwright Henrik Ibsen's (1828-1906) letters using simple latent lexical variables in order to model their degree of surprise. By temporally sorting the roughly 2,400 letters by date, we extract their novelty and resonance time series and apply Bayesian change point detection in order to detect time-dependent significant changes. We align the letters with Ibsen's literary career in order to test competing hypotheses about its development, based on qualitative studies of Ibsen, about the location of context-dependent change points in Ibsen's playwriting. Our preliminary results show that change points are likely to be located in 1873 around the publication of the historical play Emperor and Galilean and in 1894 around the time Little Eyolf was published. We then apply the same approach to summaries of all of Ibsen's plays and evaluate the bi-directional predictive relationship between the letters and the plays. We discuss these findings in relation to ingrained literary historical narratives about ostensible watershed moments in Ibsen's career, which typically posit change points with the publication of Brand in 1866 and A Doll's House in 1879. Other Ibsen scholars locate a third change point in 1894, which corresponds with our preliminary findings. This study has potential literary historical significance because Ibsen's career is closely associated with entrenched literary categories and change points, such as the rise of "Realism" and the so-called "Modern Breakthrough" in Scandinavian literature. Testing change in Ibsen's career may provide evidence for an alternate narrative that is less agonistic, and more indicative of continuity over time.

References

1 A. T. Barron, J. Huang, R. L. Spang, S. DeDeo, Individuals, institutions, and innovation in the debates of the french revolution, Proceedings of the National Academy of Sciences 115 (2018) 4607–4612.

2 K. L. Nielbo, F. Haestrup, K. C. Enevoldsen, P. B. Vahlstrup, R. B. Baglini, A. Roepstorff, When no news is bad news–detection of negative events from news media content, arXiv preprint arXiv:2102.06505 (2021).

3 M. Wevers, J. Kostkan, K. L. Nielbo, Event flow - how events shaped the flow of the news, 1950-1995, in: CHR, 2021.

4 K. L. Nielbo, K. F. Baunvig, B. Liu, J. Gao, A curious case of entropic decay: Persistent complexity in textual cultural heritage, Digital Scholarship in the Humanities (2018).

An OCR Pipeline for Transforming Parliamentary Debates into Linked Data: Case ParliamentSampo – Parliament of Finland on the Semantic Web

Senka Drobac[1], Laura Sinikallio[2, 1], Eero Hyvönen[1, 2]

[1]Aalto University, Finland; [2]University of Helsinki, Finland

Parliamentary data are used in many areas of research, 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 discuss and vote on issues on the agenda and other topics that arise. Semantic Web (SW) technologies and Linked Data (LD) enable the publishing and using parliamentary data in Digital Humanities. Linked data and ontologies provide a framework for harmonizing heterogeneous distributed datasets and combining them into larger and richer entities. Moreover, when the machine "understands" the content of the data, intelligent web services and data analyses can be implemented more easily. Finally, ready-made tools by other actors can be reused for publishing, processing and analyzing the standardized data.

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 RDF4. 4) The data can be enriched, interlinked and finally published 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. Metadata extraction and enrichment (steps 3–4) are described in [2, 3, 4].

The parliamentary sessions from the period 1907-1999 have been scanned and made available as PDF files. We OCRed the data with Tesseract 4, with a combination of the pre-trained Finnish and Swedish models. Due to the large size of the dataset (324,333 page images), we had to perform the recognition using supercomputers. That enabled us to process the entire dataset in a reasonable time.

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. To enhance the reliability of the gathered data, we performed a few manual corrections to the OCR results. After corrections, we created Python scripts to scrape all relevant data from the OCRed text files. Finally, we transformed the speeches into two parallel data sets: (1) an RDF (Resource Description Framework) format speech knowledge graph, forming linked data and (2) an XML corpus formed according to the Parla-CLARIN v0.2 specification.[5]

The speech data outcome described in this paper has already been used as a basis for analyzing concepts in political speeches, [6] for network analyses based on MP references in speeches, [7] for data analyses of speeches and for portal application development.[4]

References

1 Hitzler, P., Krötzsch, M., Rudolph, S.: Foundations of Semantic Web technologies. Springer (2010).

2 Sinikallio, L., Drobac, S., Tamper, M., Leal, R., Koho, M., Tuominen, J., Mela, M.L., Hyvönen, E.: 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. pp. 1–17. Schloss Dagstuhl- Leibniz-Zentrum fur Informatik GmbH, Dagstuhl Publishing (August 2021).

3 Leskinen, P., Hyvönen, E., Tuominen, J.: Members of Parliament in Finland knowledge graph and its linked open data service. In: Graphs. Proceedings of the 17th International Conference on Semantic Systems, 6-9 September 2021, Amsterdam, The Netherlands. pp. 255–269 (2021).

4 Hyvönen, E., Sinikallio, L., Leskinen, P., Mela, M.L., Tuominen, J., Elo, K., Drobac, S., Koho, M., Ikkala, E., Tamper, M., Leal, R., Kes¨aniemi, J.: 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 (May 2022).

5 Erjavec, T., Pančur, A.: Parla-CLARIN - a TEI schema for corpora of parliamentary proceedings (May 2022).

6 Elo, K., Karimäki, J.: Luonnonsuojelusta ilmastopolitiikkaan: Ympäristöpoliittisenkäsitteistön muutos parlamenttipuheessa 1960–2020. Politiikka 63(4) (Nov 2021).

7 Pokkimäki, H., Leskinen, P., Tamper, M., Hyvönen, E.: Analyses of networks of politicians based on linked data: Case ParliamentSampo – Parliament of Finland on the Semantic Web (2022), paper under peer review.

Phytobibliography: Building a Digital Database of Plants in Scandinavian Picturebooks for Children

Beatrice G. Reed

Western Norway University of Applied Sciences, Norway

Despite being indispensable to all other life on earth, there is a tendency in western society to overlook plants, usually viewing them as passive backdrops of animal actions and behaviour. As the capacity of plants to sense, communicate and learn has been recognized through recent biological research (Trewavas, Mancuso and Viola), this phenomenon, labelled "plant blindness" by botanists and biology educators J. H. Wandersee and E. E. Schussler, has gained increasing attention even in the environmental humanities. With the growing acknowledgement of plants and their importance to human life and society, there is a need for developing new ways to review literature and art, including digital methods.

The paper addresses this need by presenting the ongoing process of building The Phytobibliographical Database, a historical overview of plants in Scandinavian picturebooks for children from 1900 to the present day. The database is designed as a quantitative and qualitative framework, developed to recognize, and register the flora of Scandinavian picturebooks for children and analyze their role and function. Due to the complex intermedial qualities of picturebooks, text-based methods such as corpus linguistics and text-mining techniques are not sufficient for examining plant representation relying on the interplay between texts and pictures. Additionally, most Scandinavian picturebooks are not digitalized. In the presentation, I aim to discuss these challenges and describe the strategies deployed while designing the database and collecting the corpus. As The Phytobibliographical Database is still a work in progress, ideas, and input on how to proceed are invited.

References

Mancuso, Stefano, and Alessandra Viola. Brilliant Green: the Surprising History and Science of Plant Intelligence. Island Press, 2015.

Trewavas, Anthony. Plant Behaviour and Intelligence. Oxford University Press, 2014.

Wandersee, James H., and Elisabeth E. Schussler. "Preventing Plant Blindness." The American Biology Teacher, vol. 61, no. 2, 1999, pp. 82-86.

Reconstructing MultiTorg: Archaeological Approaches to Digital Artefacts of the Historical Web

Jon Carlstedt Tønnessen
National Library of Norway, Norway

In May 1993, the very first website on the Norwegian top-level domain was published: MultiTorg. At this "electronic marketplace", visitors could read instant news, download Paul McCartney's latest hit, play video clips of upcoming Nintendo games, and view the most recent satellite image of the Nordic and Baltic countries.

MultiTorg was published at the very dawn of web media, years before the development of methods and standards for web archiving. Thus, it escaped preservation in a format that can be inspected and replayed with the use of "Wayback Machines". However, one of the creators made a security backup of the server in September 1993, providing a time-situated copy of the web server.

The case of MultiTorg raises interesting questions about how researchers can engage with early web media, produced before the standardisation of web archiving. This presentation will discuss an exploratory strategy, adapting archaeological approaches to excavate, analyse, and reconstruct digital artefacts from the past. The concept of a digital archaeology, as suggested by Dutch computer scientists Gerard Alberts, Marc Went and Robert Jansma, should not be confused with the more metaphorical use, commonly found in media archaeology or Foucauldian studies of knowledge. Rather, it aims to regard the backed-up file system as a digital field for excavation: a logically defined space with scattered fractures and remains of digital artefacts from the past that once formed a whole.

The presentation will show the process of excavation, analysis and reconstruction, and discuss some main possibilities and challenges associated with the digital archaeology approach.

References

Aasman, Susan, Tjarda De Haan, and Kees Teszelszky. "Web Archaeology: An Introduction". TMG Journal for Media History 22, no. 1 (2019): 66–84.

Alberts, Gerard, Marc Went, and Robert Jansma. "Archaeology of the Amsterdam Digital City; Why Digital Data Are Dynamic and Should Be Treated Accordingly". Internet Histories 1, no. 1–2 (2017): 146–59.

Brügger, Niels, and Ian Milligan (eds). The SAGE Handbook of Web History. Los Angeles: SAGE, 2019.

Brügger, Niels, and Ralph Schroeder. Web as History: Using Web Archives to Understand the Past and the Present. London: UCL Press, 2017.

De Haan, Tjarda. "Project 'The Digital City Revives' A Case Study of Web Archaeology." iPRES2016, 16th International Conference on Digital Preservation, 2016.

De Haan, Tjarda, and Erwin Verbruggen. "Webarcheologie, Schatgraven in En Bewaren van Het Recente (Born-)Digital Verleden: Een Praktische Handleiding". TMG Journal for Media History 22, no. 1 (2019): 1–5.

Jansma, Robert. "Metadata Dating the Digital City: A Software Archaeological Approach". Internet Histories 6, no. 4 (2022): 1–21.

Jansma, Robert. "Scoops and Brushes for Software Archaeology: Metadata Dating". University of Amsterdam, 2020.

Knijff, Johan van der. “Recovering ‘90s Data Tapes: Experiences From the KB Web Archaeology Project.” Amsterdam, 2019.

Knijff, Johan van der. Resurrecting the 1st Dutch web index: NL-menu revisited, 2018.

Lemonnier, Pierre. Elements for an Anthropology of Technology. Anthropological Papers 88. University of Michigan, 1992.

Milligan, Ian. The Transformation of Historical Research in the Digital Age. Elements in Historical Theory and Practice. Cambridge: Cambridge University Press, 2022.

Nanni, Federico. “Reconstructing a Website’s Lost Past: Methodological Issues Concerning the History of Unibo.It”. Digital Humanities Quarterly 11, no. 2 (2017).

Teszelszky, Kees. “The Historic Context of Web Archiving and the Web Archive: Reconstructing and Saving the Dutch National Web Using Historical Methods”. In The Historical Web and Digital Humanities, 13–28. Routledge, 2019.

Teszelszky, Kees. “Web Archaeology in The Netherlands: The Selection and Harvest of the Dutch Web Incunables of Provider Euronet (1994–2000)”. Internet Histories 3, no. 2 (2019): 180–94.

Webster, Peter. “Digital Archaeology in the Web of Links: Reconstructing a Late-1990s Web Sphere”. In The Past Web: Exploring Web Archives, edited by Daniel Gomes, Elena Demidova, Jane Winters, and Thomas Risse, 155–64. Cham: Springer International Publishing, 2021.

Search and Exploring Correspondence: correspSearch

Stefan Dumont, Sascha Grabsch, Ruth Sander
Berlin-Brandenburg Academy of Sciences and Humanities, Germany

Letters, not least due to their heterogeneous contents, are an intriguing and valuable source of research information all across the humanities. Despite the multitude of research applications, the sheer mass of available correspondences causes many to remain unexamined and uncited (Bunzel). In this show-and-tell, we wish to present our solution: the web service correspSearch, developed by the Berlin-Brandenburg Academy of Sciences and Humanities. The presentation will provide a brief overview of the service and a short demonstration of its functions.

Developed to aggregate letter metadata, correspSearch takes steps towards fulfilling the community's call for an extensible, decentralized, open, and digital platform that associates existing letters with one another (Bunzel). The web service predominantly uses XML technologies and operates with minimal TEI standards and linked data.

Conceptually open regarding the subject, time period, and region, correspSearch strives to help scholars search for edited letters and guide them to the original publications. Thereby the service provides researchers with a tool to explore a greater network of letters. Various filter options, such as by person, date, edition, and profession, provide further modes of exploration. correspSearch aggregates its data from digital indices of letters, which follow a specific TEI XML-based interchange format (CMIF). Once an institution provides correspSearch with CMIF files of their letter collection via their own URL, correspSearch retrieves the metadata in periodic intervals. This allows for easy expansion and updating of the indices. The aggregated data is searchable via our website as well as queryable via an API. All data is of course accessible under free licenses for further reuse.

correspSearch relies on the support of the community for metadata of various letter collections to expand its service. As of October 2022, correspSearch has collected data of over 180,000 letters from over 300 publications, and the numbers continue to rise.

References

Bunzel, Wolfgang. 2013. "Briefnetzwerke der Romantik. Theorie – Praxis – Edition." In Brief-Edition im digitalen Zeitalter, edited by Anne Bohnenkamp and Elke Richter, 109–131. Beihefte zu editio, vol. 34. Berlin: de Gruyter.

Dumont, Stefan. 2016. "correspSearch – Connecting Scholarly Editions of Letters". Journal of the Text Encoding Initiative (10).

Dumont, Stefan, Grabsch, Sascha, Müller-Laackman, Jonas, Sander, Ruth: correspSearch - connect scholarly editions of correspondence (2.0.0) [web service]. Berlin-Brandenburg Academy of Sciences and Humanities (2021).

Some Nobel Laureates Are More Coherent Than Others: Measuring Literary Quality in a Corpus of High Prestige Contemporary Literature

Yuri Bizzoni[1], Pascale Feldkamp Moreira[2], Ida Marie Lassen[1], Kristoffer Laigaard Nielbo[1], Mads Rosendahl Thomsen[1]

[1]Aarhus University, Denmark; [2]Utrecht University, The Netherlands

The definition of literary quality is a highly complex problem that different schools of thought have approached through very different lenses. Studies that link literary quality to stylometric properties of the texts often look at global features, such as average sentence length,[1] but recent studies show that a text's development can be a better predictor of readers' appreciation.[2] Hypothesising that a balance between coherence and novelty improves the reading experience,[3] examined the entropy of features linearly across literary texts, while[4] looked at the fractal dynamics of their sentiment arcs.

Recently,[5] used sentiment arcs' coherence to tell Nobel laureates' works from control groups, based on the idea that the Nobel Prize is the ultimate recognition of literary quality. However, Nobel laureates' production can vary, and some laureates' text can be more appreciated than others.

In our work, we use GoodReads' scores to find out whether metrics that correlate with quality in broader corpora keep their predictive power within the Nobel canon itself. Defining the coherence of sentiment arcs through three complementary measures - fractality,[4] approximate entropy[3] and arc compressibility[6] - we find that novels within certain intervals of these measures tend to elicit higher ratings, suggesting that they might capture a balance between predictability and novelty which remains powerful even among highly competitive writers.

References

1 A. van Cranenburgh and R. Bod, "A Data-Oriented Model of Literary Language", in Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, vol. 1, 2017, pp. 1228–1238.

2 S. Maharjan, S. Kar, M. Montes-y-Gomez, F. A. Gonzalez, and T. Solorio, "Letting Emotions Flow", Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics, vol. 2., 2018, pp. 251-265.

3 M. Mohseni, C. Redies, and V. Gast, "Approximate Entropy in Canonical and Non-Canonical Fiction", Entropy, vol. 24, no. 2, p. 278-294, 2022.

4 Y. Bizzoni, T. Peura, M. R. Thomsen, and K. Nielbo, "Sentiment Dynamics of Success: Fractal Scaling of Story Arcs Predicts Reader Preferences", NLP4DH, 2021, pp. 1-6.

5 Y. Bizzoni, T. Peura, M. R. Thomsen, and K. Nielbo, "The Fractality of Sentiment Arcs for Literary Quality Assessment: The Case of Nobel Laureates". NLP4DH, in print 2022.

6 C. Koolen, K. van Dalen-Oskam, A. van Cranenburgh, and E. Nagelhout, "Literary Quality in the Eye of the Dutch Reader: The National Reader Survey", Poetics, vol. 79, 2020, pp. 1-13.

Towards Reusable Aggregated Biographical Research Data: Provenance and Versioning in the InTaVia Knowledge Graph

Joonas Kesäniemi[1], Matthias Schlögl[2], Jouni Tuominen[1,3], Victor de Boer[4], Go Sugimoto[4]

[1]Aalto University, Finland; [2]Austrian Academy of Sciences, Austria; [3]University of Helsinki, Finland; [4]Vrije Universiteit Amsterdam, The Netherlands

InTaVia is an EU Horizon 2020 funded project with an aim to provide researchers and the informed public access to a) a large Knowledge Graph containing heterogeneous multilingual data from Austria, Finland, Slovenia, and the Netherlands from national biographies and b) an easy to use a web-based tool to analyze this data with visual analytics (VA) methods. Robust versioning and provenance are paramount because although some of the biography datasets are static, their digital representations and enriched data changes as technologies, methods and disciplines evolve.

As the InTaVia Knowledge Graph (IKG) matures and the amount of data increases, challenges related to the scalability and usability of both versioning and provenance become more prominent. Also, keeping query times low and providing users with provenance information that they can actually use becomes more complicated (Sikos e al. 2020).

To solve these problems, the InTaVia platform implements a layered approach to provenance management. The IKG consists of a set of named graphs that are maintained and versioned exclusively through manually and automatically executed workflows. This allows for a complete processing level provenance (e.g. execution date and ID of the workflow run) to be collected with regard to input and target datasets. Workflows also follow shared implementation conventions and they must provide descriptions of the individual target datasets they produce. The other layer of provenance is part of the data graph and related to the research objects themselves, such as the data source and biographer.

Versioned named graphs are accessible to the user through a Memento (Sompel et al. 2009) inspired API that provides users with a consistent view of the IKG even if the underlying data changes. It also enables applications using the API to notify the user if there is an updated version of the data available.

References

Sikos, L.F., Philp, D. Provenance-Aware Knowledge Representation: A Survey of Data Models and Contextualized Knowledge Graphs. Data Sci. Eng. 5, 293–316 (2020).

Van de Sompel, H., Nelson, M., Sanderson, R., Balakireva, L.Michael L., Ainsworth, S., Shankar H. Memento: Time Travel for the Web. arXiv.org (2009).

Transforming Linguistically Annotated Finnish Parliamentary Debates Into the Parla-CLARIN Format

Minna Tamper[1,2], Laura Sinikallio[1], Jouni Tuominen[1], Eero Hyvönen[1,2]
[1]University of Helsinki, Finland; [2]Aalto University, Finland

The debates in the Parliament of Finland take place in public plenary sessions whose minutes have been transcribed and published in different formats, as printed minutes, as HTML pages, and using a custom XML format, since the Parliament convened for the first time in 1907. The Semantic Parliament project with its data services and the ParliamentSampo portal makes all ca.1 million Finnish parliamentary speeches 1907-2022 accessible for studying parliamentary politics, culture, and language.[1,2] The data has been transformed into linked open data and also into the TEI-based specialization of the Parla-CLARIN format in the ParlaMint multi-national project[3] that contributes to the creation of comparable and uniformly annotated multilingual corpora of parliamentary sessions. This paper presents the work done to transform the Finnish parliamentary debates from 2015 to the present day into the Parla-CLARIN format in ParlaMint. The new dataset includes not only the transcripts for the debates[1] but also the linguistically annotated corpus with named entities extracted from the texts.

References

1 Laura Sinikallio et al.: Plenary Debates of the Parliament of Finland as Linked Open Data and in Parla-CLARIN Markup. 3rd Conference on Language, Data and Knowledge, LDK 2021, Open Access Series in Informatics (OASIcs), vol. 93, pp. 8:1-8:17, August, 2021.

2 Eero Hyvönen et al.: Finnish Parliament on the Semantic Web: Using ParliamentSampo Data Service and Semantic Portal for Studying Political Culture and Language. Digital Parliamentary data in Action (DiPaDa 2022), Workshop at the 6th Digital Humanities in Nordic and Baltic Countries Conference, CEUR Workshop Proceedings, Vol. 3133, May, 2022.

3 Tomaž Erjavec et al., ParlaMint: Comparable Corpora of European Parliamentary Data. In: (M. Monachini and M. Eskevich, eds.) Proceedings of CLARIN Annual Conference 2021. 2021, pp. 20-25.

Virtual Lab at the National Library of Estonia

Peeter Tinits[1], Urmas Sinisalu[2]
[1]University of Tartu, Estonia; [2]National Library of Estonia, Estonia

Cultural heritage institutions have amassed large digital collections that are starting to be explored in new ways – as datasets to analyze and explore. The GLAM labs initiative has sought to bring these resources closer to researchers following Open Science principles with the help of public code interfaces, linked and findable datasets.

The National Library of Estonia is building a Virtual Lab to facilitate data-based access and use for its local collections. The Virtual Lab will feature digital collections made into datasets as well as a text access point. It will collect and showcase case studies built upon them to demonstrate the value found in the use of the collections. The initiative is supported by the Open Digital Libraries project, the Lab website is due to launch in February 2022.

We will present the journey there and the current stage of the work. In setting up the lab, we performed a legal analysis and a service design study, conducting interviews with local key figures and potential typical users. We will present the results and some implications from these studies. In collaboration with the University of Tartu (project EKKD72), we also ran a pilot project in building case studies on the library newspaper collections, collecting feedback and developing tools in the process.

The Virtual Lab aims to offer data and resources to researchers of diverse backgrounds as well as providing an outlet to present their work and integrate the results and new data into library workflows. Through the Lab, we seek to establish a greater partnership between libraries and educational institutions that would allow both the data stewards and data users to gain the most out of easier access to digital collections.

References

EKKD72 - The use of textual materials in digital humanities case studies on the example of Estonian newspaper collections (1850-2020). (01.01.2022-31.12.2022)

Open Digital Libraries - A Creative Europe project on generating new value out of library digital collections through creative reuse. (01.09.2020-31.08.2023)

GLAM Labs - An international network of galleries, libraries, archives and museums that seeks to foster the creative reuse of digital collections in their institutions.

Visualising the Cuneiform Corpus: Results of the Project Geomapping Landscapes of Writing (GLoW)

Seraina Nett[1], Nils Melin-Kronsell[1], Carolin Johansson[1], Gustav Ryberg Smidt[2], Rune Rattenborg[1]

[1]Uppsala University, Sweden; [2]Ghent University, Belgium

Counting upwards of half a million documents, the corpus of cuneiform texts, written in cuneiform script on clay tablets, is one of the largest bodies of written sources from the ancient world. Cuneiform texts encompass a diverse range of genres, from economic documents to literary and scholarly texts, and are found across most of the Middle East and in areas as far afield as Egypt, Afghanistan, and Italy. The texts span a chronological range from c. 3,200 BCE to the 1st century CE and are composed of a range of languages that are part of different language families. Because of the size and extreme temporal and spatial spread of the corpus, no attempt has been made so far to map and analyse this corpus in full.

The three-year research project Geomapping Landscapes of Writing (GLoW), based at Uppsala University and generously funded by Riksbankens Jubileumsfond, aims to produce an updated, global survey of cuneiform inscriptions in collaboration with existing digital text catalogues and open access data repositories, such as the Cuneiform Digital Library Initiative (CDLI). Using GIS-aided spatial analysis and other digital humanities research tools, the project explores our newfound technological ability to produce a reliable overview of this immense corpus and to analyse the distribution of this corpus in meaningful ways and aims to further document this diverse corpus of cultural heritage dispersed in museums around the world.

In this show-and-tell presentation, we will provide an updated overview of the cuneiform corpus as a cultural heritage collection based on the metadata generated by the GLoW project, thus exploring and visualising the distribution of the corpus in terms of categories such as geography, genre, language and others.

References

Bigot Juloux, V., A.R. Gansell, and A. Di Ludovico (2018). CyberResearch on the Ancient Near East and Neighboring Regions. Leiden & Boston: Brill.

Harrison, T.P. (2018). "Computational research on the Ancient Near East (CRANE): large-scale data integration and analysis in Near Eastern archaeology". Levant 52:1-2, 1–4.

Pedersén, O. (2012). "Ancient Near East on Google Earth: Problems, Preliminary Results, and Prospects". In: Proceedings of the 7th International Congress on the Archaeology of the Ancient Near East 12 April - 16 April 2010, the British Museum and UCL, London. Vol. 3: Fieldwork & Recent Research. Ed. by R. Matthews and J. Curtis. Wiesbaden: Harrassowitz, 385–393.

Rattenborg, R., C. Johansson, S. Nett, G. R. Smidt, J. Andersson 2021a: "An Open Access Index for the Geographical Distribution of the Cuneiform Corpus". Cuneiform Digital Library Journal 2021 (1): 1–12.

———: 2021b. "Cuneiform Inscriptions Geographical Site Index (CIGS)". Zenodo [data set].

Streck, M. P., 2010: "Großes Fach Altorientalistik: Der Umfang des keilschriftlichen Textkorpus". Mitteilungen der Deutschen Orient-Gesellschaft 142: 35–58.

The Words of Climate Change: TF-IDF-Based Word Clouds Derived From Climate Change Reports

Maria Skeppstedt[1], Magnus Ahltorp[2]

[1]Uppsala University, Sweden; [2]Swedish Institute for Language and Folklore, Sweden

Traditional word clouds have received criticism. Still, easily interpretable visualisations of long texts can be useful, e.g. for raising interest in the text content. We aim to construct such visual overviews of climate change reports. Building on previous research on further developments of the traditional word cloud,[1, 2, 3] we constructed word clouds with the following non-standard properties: (i) Word and bigram prominences are determined by TF-IDF (term frequency-inverse document frequency) instead of frequency, (ii) prominence is not only indicated by font size, but also by the vertical position of the words/bigrams and by bars with a height proportional to their TF-IDF values, (iii) the horizontal position of the words/bigrams is determined by a one-dimensional t-SNE visualisation of word2vec-vectors.

Seven different reports were visualised, five Swedish translations of IPCC-report summaries for policymakers, one report on thought structures that hinder climate change mitigation, and one report from the Swedish Climate Policy Council. We extracted and visualised the 500 most prominent words/bigrams for each report.

Apart from potentially raising interest in the report content, the visualisations also provide an overview of the type of vocabulary required for understanding the reports. We, therefore, aim to investigate whether the word clouds also might form an inspiration when constructing dictionaries and other language resources for the climate change domain.

The open-source code for generating the word clouds can be found at CDH Uppsala word-rain.

References

1 Lukas Barth, Stephen G. Kobourov, & Sergey Pupyrev. 2014. Experimental comparison of semantic word clouds. In Joachim Gudmundsson & Jyrki Katajainen, editors, Experimental Algorithms, pages 247–258, Cham. Springer International Publishing.

2 Erich Schubert, Andreas Spitz, Michael Weiler, Johanna Geiß, & Michael Gertz. 2017. Semantic word clouds with background corpus normalization and t-distributed stochastic neighbor embedding, ArXiv.

3 Fernanda B. Viégas & Martin Wattenberg. 2008. Timelines tag clouds and the case for vernacular visualization. Interactions, 15(4):4952, jul.

Workshops

Challenges of Long-Term Sustainability of DH Projects: The LAM (Libraries, Archives and Museums) Perspective

Olga Hołownia[1], Helena Byrne[2], Grace Bicho[3]

[1]International Internet Preservation Consortium, United States of America; [2]The British Library, Great Britain; [3]The Library of Congress, United States of America

Status of the Field

7 Mar 2023
15:00-17:00

In 2009, an entire issue of Digital Humanities Quarterly was dedicated to a topic that is likely to resonate with anyone ever involved in managing a Digital Humanities project – "How do we know when we're done? What does it mean to "finish" a piece of digital work?" (Kirschenbaum, 2009). The open-ended nature of DH projects can be quite liberating but may also result in limited planning for long-term sustainability, discoverability and access to completed projects. Funding focused on starting up does not help with creating standards and workflows for the preservation of digital projects. What happens to project websites and databases once the funding runs out? What is the "natural" home point once the projects have been completed? Regarding lessons learned, does it remain important to ask; "Whatever happened to Project Bamboo"? Can web archives support long-term preservation and access? A number of issues related to the longevity of DH projects have been more recently discussed in the context of King's Digital Labs and lessons learned from their experiment are a good starting point for this workshop along with the initiative by the Portuguese Web Archive to preserve research and development project websites hosted on the .eu domain. How far do current practices in research data management and the development data management plans cater for the long-term preservation requirements of DH projects? Could ALM institutions partner with research infrastructures such as DARIAH and CLARIN to address this issue?

While partnerships with cultural heritage institutions won't necessarily mean that they will become long-term custodians of outputs of DH projects, their expertise can certainly contribute significantly to the development of best practices. Increasingly, university IT departments are introducing policies that define shutdown dates for project websites and databases that have not been regularly maintained, and are viewed as security risks. Due to playback and access limitations, web archives are still not considered a good solution for the project's ultimate "shelf life". Yet, to date, there is no long-term solution to preserve the institutional memory of these, often publicly funded, DH initiatives.

Aims of the Workshop

The aims of this workshop are: 1) to explore current practices related to long-term preservation of DH projects, including websites and custom-built databases, 2) to identify challenges related to maintenance as well as storage and cataloging of such initiatives, 3) identify barriers of archiving and accessing such content via web archives, and 4) to work out shared solutions for providing discoverability and access to such projects. We would also look at the ways ALM institutions help promote DH

projects, e.g. through educational programmes, GLAM Labs initiatives and linking the projects to existing digital collections, etc.

Output

The workshop aims to bring together DH researchers and ALM practitioners interested in the topic and would focus on current standards and workflows, ways of enhancing discoverability, use cases as well as outreach programmes. Invited speakers will present short talks on best practices in their institutions. This will be followed by a brainstorming breakout session the goal of which is to create generic guidelines that could be used to create workflows for individual institutions and web domains where the projects are hosted. This workshop follow-up on the DHNB 2022 workshop Digital Humanities and Support Units in the Nordic and Baltic Countries (2022) could contribute to connecting the activities of the DHLAM Working Group and the IIPC Research Working Group.

References

Dombrowski, Q. (2014). What Ever Happened to Project Bamboo? Literary and Linguistic Computing, 29(3).

Kirschenbaum, M. G. (2009). Done: Finishing projects in the Digital Humanities. Digital Humanities Quarterly, 3(2).

Smithies, J. et al (2019). Managing 100 Digital Humanities Projects: Digital Scholarship & Archiving in King's Digital Lab. Digital Humanities Quarterly, 13(1).

H2020 projects preserved by Arquivo.pt (2021).

Cross-University Collaboration in Digital Humanities and Social Science (DHSS) & Digital Humanities & Cultural Heritage (DHCH) Education

Jonas Ingvarsson[1], Ahmad Kamal[2], Koraljka Golub[2], Isto Huvila[3], Olle Sköld[3], Anna Foka[3], Marianne Ping Huang[4], Mikko Tolonen[5]

[1]University of Gothenburg, Sweden; [2]Linnaeus University, Sweden; [3]Uppsala University, Sweden; [4]Aarhus University, Denmark; [5]University of Helsinki, Finland

Workshop Description

7 Mar 2023
08:30-13:30

The workshop will allow established and recent DH programs and educational initiatives to report on their experiences, allowing us to form a diverse community of pedagogues to critically discuss topical issues relating to DH education. The DHNB venue encourages participation by teachers, researchers and developers from different perspectives. As the sixth workshop on education at DHNB, this year will focus on higher education in DH, aiming at pedagogical development and infrastructure building. Workshop participants will share their DH teaching experiences, including discussions of strategies, tools, platforms, evaluations, outcomes, and problems. The workshop will also explore a series of initiatives from the DHNB Higher Education Working Group intended to enable collaborative education among fellow universities with DHSS/DHCH programmes/courses/modules. Currently, there are three primary initiatives of the Higher Education Working Group, which we will address in the workshop's main session: (1) DHSS/DHCH course exchanges, (2) project-based education, and (3) DHSS/DHCH instruction seminars. Visit the workshop webpage at Linnaeus University.

Workshop Themes

- Collaborations/exchanges in digital humanities (DH) instruction
- Project-based/problem-based DH education
- Interdisciplinary/cross-disciplinary/cross-sectoral/international cooperation in DH education
- Existing programs, modules or individual courses in DH (e.g., design, target student groups, content, job market, evaluation, experiences, lessons-learned)
- Currently developed programs, modules or individual courses in DH (e.g., design choices, target student groups, resource management, related issues)
- Capacity building for student employability

Programme

8:30 - 8:40 Welcome and introductions
8:40 - 9:30 Panel 1: Teaching Experiences I

- An open educational resource for teaching digital humanities skills: The cultural analytics open science guide – Federico Pianzola (University of Groningen)
- Global, social and cultural competencies of future EFL teachers: Germany-Ukraine universities cooperation – Maria Eisenmann (University of Wuerzburg), Anatoliy Prykhodko (Zaporizhzhia Polytechnic National University), Nataliia Lazebna (University of Wuerzburg), & Kateryna Lut (University of Wuerzburg)
- Let's tweet again: Twitter as a tool for master students – Elena Duce Pastor (Autonoma University of Madrid)

9:30 - 9:40 *Coffee break*
9:40 - 10:30 Panel 2: Teaching Experiences II

- Programming and data visualization for academic audiences across institutions and disciplines: Lessons learned – Andres Karjus (Tallinn University; Datafigure Plc.)
- rp4if.Teaching IIIF on Raspberry Pis – Wout Dillen (University of Borås) & Joshua Schäuble (University of Groningen)
- Research-based teaching for better language and linguistics careers – Maja Miličević Petrović (University of Bologna), Tanja Samardžić (University of Zurich), Darja Fišer (CLARIN), Silvia Bernardini (University of Bologna), Iulianna van der Lek (CLARIN), Boban Arsenijević (University of Graz), & Marko Simonović (University of Graz)

10:30 - 10:40 *Coffee break*
10:40 - 11:30 Panel 3: Project-Based Learning

- Experimentation in project-based education in DH – Ernesta Kazakėnaitė (Vilnius University) & Justina Mandravickaitė (Vilnius University)
- Engaging students in digital humanities project of digitization, cataloguing and providing open access to the Ivo Maroević's slide collection – Goran Zlodi (University of Zagreb)
- Project-based approach to digital humanities in university education – Bence Vida Tivadar (ELTE University) & Palkó Gábor (ELTE University)

11:30 - 11:40 *Coffee break*
11:40 - 12:30 Nordic DH Education Updates

- Training in the Swedish national infrastructure for humanities – Coppelie Cocq (Umeå University), Koraljka Golub (Linnaeus Unviersity), Marianne Gulberg (Lund University) & Cecilia Lindhé (Gothenburg University)
- DASH: A PhD network for DH students in Sweden – Anna Foka (Uppsala University)
- DH Reports from Finland – Mikko Tolonen (Helsinki University)
- Revising programming instruction for DH students – Ahmad Kamal (Linnaeus University)

12:25 - 12:35 *Coffee break*
12:35 - 13:25 Working Group Initiatives

- Reporting on the DH student exchange survey – Jonas Ingvarsson (University of Gothenburg) & Ahmad Kamal (Linnaeus University)
- Further discussions: Project-based learning support, teaching workshops, and other potential initiatives 13:25-13:30 Concluding remarks & action plans

---

Exploring Digital Tools and Platforms for Individual Research of History and Antiquity

Victoria G. D. Landau, Sarah Siegenthaler
University of Basel, Switzerland

Exploring Digital Tools and Platforms for Individual Research of History and Antiquity

9 Mar 2023
09:00-12:00

Digital history, computational archaeology, digital heritage — the digitization of humanities research has inevitably "arrived" at the doorstep of disciplines dealing with the past, both recent and long gone. And though its approach has been gradual, many institutions have unfortunately displayed how ill-equipped they remain to address it. With ancient civilizations departments often lacking funds for even their regular curriculum, DH-oriented courses for their undergraduate and graduate students, or even more practical tutorials for their postgraduate researchers, remain scant. While institutions may have forcibly remained stagnant, scholars themselves have stepped up and adapted to an ever-changing research landscape and job market. This has resulted in the phenomenon of self-trained, self-tested "amateur" users, with trained scholars ending up feeling like non-experts in tools they regularly (and competently) use.

Aim of the Workshop

This workshop is conceptualized as an informal scientific exchange on precisely these digital helpers and infrastructures we use both every day and for specific needs, where no tool is too rudimentary to work with as long as it is suitable – everyone is aware of the great potential and silent power of a simple collaboration platform like the online "Google Suite" (Docs, Sheets, Slides, Forms et al.), but also the degree of competency one can achieve with a system like QGIS. Seeing this as a broad exhibition of available tools, we will span open-source tools and platforms, data management structures, universal implementations (e.g. IIIF) all the way to actual (geographic) mapping and visualization of research results, connections and networks.

Tentative Programme

Preliminarily, we will structure the half-day into the sections "Doing/During research" and "Presenting research", but of course these two sections bleed into one another depending on the tool, a point which we will address and discuss during the workshop. The participating scholars will be contributing as much insight to the workshop as the organizers, who aim to provide a receptacle for exchange and community among researchers facing many of the same issues across disciplines/subjects and who have developed their own unique solutions using digital tools. The workshop as such is therefore not just intended for historians, archaeologists, philologists and the like, but the case studies we will look at will derive from and focus on ancient civilizations. Participants from all backgrounds and prior experience are welcome to

join. Participants signing up will also be asked to submit one or more examples for tools they currently utilize or have previously used in their research, to get an idea of their experiences and to include in the plenary parts of the workshop. Additionally, they can pose a question on the topic in advance which will either be addressed individually or be discussed amongst participants to exchange tips and opinions. All of the tools/platforms presented and collected throughout the workshop will be assembled and made available online (e.g. Github, Notion) after the DHNB2023 conference, to properly harness the results of the exchange.

KUB Datalab's Digital Humanities Workshop

Lars Kjær
The Royal Danish Library, Denmark

7 Mar 2023
10:00-12:00

To mark the start of the DHNB2023 online conference, the KUB Datalab will organize an on-site activity at Copenhagen University Library, South Campus. The organization will be in collaboration with a local DH network, for example, DH-Cult.

Students and researchers at South Campus are invited to join our workshop, which will shed light on what digital humanities is. There will be four short lectures that explain how digital humanities is part of both education and research at Copenhagen University, South Campus.

Programme:

10:00 - 10:15 Welcome, presentation of the program, the DHNB, and DH activities at South Campus by Lars Kjær

10:15 – 10:35 Digital methods in Chinese – illuminated by a research of dating profiles by Anna Davidsen Buhl

10:35 – 10:55 Digital methods in History – illuminated by a research of Danish industrial development by Mikkel Støvring Hansen

10:55 – 11:05 *Short Break*

11:05 – 11:20 Experiences from the classrooms and the introduction of digital methods in the teaching at Department of English, Germanic and Romance Studies by Robert Rix

11:20 – 11:40 Experiences from the classrooms and the introduction of digital methods in the teaching at Department of Cross-Cultural and Regional Studies by Bo Ærenlund Sørensen

11:40 – 11:50 Presentation of the KUB Datalab Calendar by Lars Kjær

11:50 – 12:00 Closing and networking

The Norwegian Web Archive: Searching and Examining the Web of the Past

Jon Carlstedt Tønnessen
National Library of Norway, Norway

6 Mar 2023
9:00-12:30

Workshop Description

The National Library of Norway and the Digital Scholarship Centre at the University of Oslo Library welcome researchers to a participatory workshop on how to search and examine historical resources in the Norwegian Web Archive (NWA). The National Library will present three different prototype services for locating and representing resources from the web archive. The participators will be able to test the prototypes themselves, working with practical exercises to inspect various parts of the NWA collection, spanning back to 2001, including resources from political parties, news publications and the public sector. Tutorials will be given, demonstrating the prototypes, before participants can practice by themselves. We will provide suggestions for practical exercises, but participants are welcome to explore the services based on their own research interests.

Aim of the Workshop

The first service to be demonstrated is a URL-based search engine, allowing researchers to locate resources from specific web domains. This engine can also be used to search for specific media types (mime types), such as text, images, and audio, with different possibilities for filtration. Second, we will present a full-text search engine which allows searching for words and terms in the web archive. Queries work with an index, based on natural language extracted from HTML-files in the archive. Last, participants can view archived resources in a "Wayback Machine" – a replay service where the archived elements of historical websites can be assembled and reconstructed.

Tentative Programme

09:00 - 09:15 Introduction to the Norwegian Web Archive
09:15-09:30 Introducing the URL and media type search engine
09:30-10:00 Practical exercise
10:15-10:30 Introducing the full-text search
10:30-10:50 Practical exercise
11:00-11:30 Pitfalls and critical examination of Wayback Machines
11:30-11:50 Practical exercise
12:00-12:30 Debrief: Experiences and discussion

To Serve Them All – Web Accessibility in Digital Humanities

Tone Merete Bruvik
University of Bergen, Norway

To Serve Them All - Web Accessibility in Digital Humanities

6 Mar 2023
09:00-12:00

Requirements for universal design of websites have been a Norwegian regulation since 2014, last revised on 1 February 2022 with a one-year implementation period. The regulation refers to the EU's Web Accessibility Directive (WAD) and the standard Web Content Accessibility Guidelines (WCAG 2.1), an ISO standard. There is similar legislation within the EU.

But do we fulfill these requirements? At the University of Bergen Library, we have more than twenty web services running, including the two Norwegian standard dictionaries Bokmålsordboka and Nynorskordboka with a broad user group from pupils to professors; a special service for PhD theses; and old services as the Norwegian Newspaper Corpus which we only support because no other institutions do.

How will it be possible for us to make sure these services are available for everyone? For example, the linguistic researchers who are visually impaired, the primary school pupil with dyslexia and the student sitting in full sunshine in a café working on a PhD thesis?

Aim of the Workshop

The purpose of the workshop is to raise the attention in the field of digital humanities that accessibility issues are essential for the outreach of our research when we create websites, applications, PDFs, or write papers.

Websites developed for research projects often have a specific user group in mind, like students or researchers. They are, in many respects, thought of as a very well-skilled group, and accessibility issues might not be the top priority. However, researchers and students are like anyone else in their need to have user-friendly tools, and for some, it is essential to take these issues into account to be able to participate in academic life.

Programme

09:00 - 09:10 Welcome
09:10 - 09:50 Introduction to Web Accessibility
09:50 - 10:00 *Short break*
10:00 - 10:20 Introductions to tools and techniques for testing
10:20 - 10:50 The participants try one or two tools on a website of their choosing
10:50 - 11:00 *Short break*
11:00 - 11:45 Discussion: Looking at cases, problems, and solutions.
11:45 - 12:00 Wrap-up

Indices

Author Index

Aadland
  Emma Josefin Ölander, 34
Achmann
  Michael, 69
Ahltorp
  Magnus, 121
Andersen
  Gisle, 13
  Trygve, 9
Andersson
  Lisa, 48
Arthur
  Paul, 83
Azar
  Michael, 26

Backman
  Agnieszka, 95
Baklāne
  Anda, 85
Baunvig
  Katrine Frøkjær, 11, 77
Bicho
  Grace, 125
Bizzoni
  Yuri, 116
Bongo
  Lars Ailo, 9
Brodén
  Daniel, 26, 93
Brodén
  Daniel, 32
Bruinsma
  Bastiaan, 107
Bruvik
  Tone Merete, 134
Byrne
  Helena, 125

Cañadas
  Rafael N., 9
Capurro
  Carlotta, 18

de Boer
  Victor, 106, 117
Dombrowski
  Quinn, 95
Drobac
  Senka, 59, 110
Dumont
  Stefan, 115
Dupré
  Sven, 18
Dwenger
  Nicole, 30

Ebel
  Carla, 106
Elo
  Kimmo, 46
Enqvist
  Johanna, 59
Ernštreits
  Valts, 89

Fagerving
  Alicia, 97
Faiz Wahjoe
  Muhammad, 59
Feldkamp Moreira
  Pascale, 116
Felsing
  Ulrike, 22
Fernandez Fernandez
  Elena, 81
Fielding
  James Matthew, 24, 75
Foka
  Anna, 127
Fornaro
  Peter, 22
Fridlund
  Mats, 26, 32, 42, 93
Frischknecht
  Max, 22

Gangopadhyay
  Nivedita, 24
Gasparini
  Andrea, 65

Gheldof
Tom, 65
Gilbert
Sofie, xvi
Ginter
Filip, 46
Gius
Evelyn, 79
Golub
Koraljka, 127
Grabsch
Sascha, 115
Grève
Sebastian Sunday, 24

Hamarowski
Bartosz, 36
Hansen
Dorte Haltrup, 20
Hendriksen
Marieke, 18
Hofland
Knut, 13
Holmila
Antero, 91
Hołownia
Olga, 125
Huber
Maximilian, 57
Huvila
Isto, 48, 127
Hyvönen
Eero, 59, 110, 118
Hæstrup
Frida, 108

Ingvarsson
Jonas, 127
Israelson
Per Gunnar, 55

Jain
Ramesh, 53
Jauhiainen
Heidi, 87
Iida, 59
Tommi, 87
Johansson
Carolin, 120
Jørgensen
Finn Arne, xvi

Kāle
Maija, 44, 53
Kaiser
Jessica, 48
Kamal
Ahmad, 127
Kanoulas
Evangelos, 18
Kawabe
Sakiko, 104
Kesäniemi
Joonas, 106, 117
Kjær
Lars, 132
Koho
Mikko, 59
Kostkan
Jan, 40
Kristensen-McLachlan
Ross Deans, 30
Kristiansen
Marita, 13
Kristeröm
Olov, 15
Kruusmaa
Krister, 38

La Mela
Matti, 59
Laippala
Veronika, 46
Landau
Victoria G. D., 130
Langeloh
Jacob, 103
Lassen
Ida Marie, 116
Leskinen
Petri, 59
Liimatta
Aatu, 71
Lindemann
Pascal, 57

Lindström
Matts, 55
Liu
Ying-Hsang, 48
Lompe
Maria, 36
Ludvigsen
Louise, 9

Magin
Elisabeth Maria, 73
Marienberg-Milikowsky
Itay, 79
Martin
Benjamin G., 63
Matsson
Arild, 15
McGuire
Michael, 26
Melin-Kronsell
Nils, 120
Møldrup-Dalum
Per, 77
Müller Gjesdal
Anje, 13

Navarretta
Costanza, 20
Nelhans
Gustaf, 42
Nett
Seraina, 120
Nielbo
Kristoffer Laigaard, 11, 40, 108, 116
Nimb
Sanni, 30
Norén
Fredrik, 63

Ojala
Jari, 91
Olsson
Leif-Jöran, 93
Leif-Jöran, 32

Paloposki
Hanna-Leena, 59
Pedersen
Bjørn-Richard, 9
Perner
Mads L., 9
Pichler
Alois, 24, 75
Pierce
Rachel Laura, 67
Pikkanen
Ilona, 59
Ping Huang
Marianne, 127
Pivovarova
Lidia, 28
Provatorova
Vera, 18

Räsänen
Venla, 91
Raemy
Julien Antoine, 22
Rantala
Heikki, 59
Rasmussen
Krista Stinne Greve, 77
Rattenborg
Rune, 120
Reed
Beatrice G., 112
Rees
Ellen, 108
Rettberg
Scott, 3
Rikters
Matīss, 44
Ristilä
Anna Kristiina, 101
Rockenberger
Annika, xvi
Ryan
John Charles, 83
Yann, 28, 71

Sander
Ruth, 115
Saulespurēns
Valdis, 85
Savcisens

Germans, 81
Schlögl
Matthias, 106, 117
Schmidt
Thomas, 57
Schumacher
Mareike, 79
Sharan
Malvika, 6
Shvetsov
Nikita, 9
Siegenthaler
Sarah, 130
Sildnes
Anders, 9
Sinikallio
Laura, 110, 118
Sinisalu
Urmas, 119
Skeppstedt
Maria, 121
Sköld
Olle, 48, 127
Smidt
Gustav Ryberg, 120
Smith
Marcus, 73
Sommerseth
Hilde L., 9
Sugimoto
Go, 106, 117
Sverdljuk
Jana, 107
Sviķe
Silga, 105
Swanstrom
Lisa, 4
Säily
Tanja, 71

Tamper
Minna, 118
Tarkka
Otto, 46
Thomsen
Mads Rosendahl, 116
Tiemann
Juliane Marie-Thérèse, xvi
Tinits
Peeter, 51, 119
Toivanen
Ida, Lindroos
Jari, 91
Tolonen
Mikko, 28, 71, 127
Tuominen
Jouni, 59, 106, 117, 118
Tønnessen
Jon Carlstedt, 113, 133

Välisalo
Tanja, 91
Vad
Kirsten, 77

Wang
Ruilin, 28
Wevers
Melvin, 40
Wolff
Christian, 69

Zeldenrust
Douwe Arjen, 61

Ängsal
Magnus P., 32, 93
Öhberg
Patrik, 32, 93

Šķirmante
Karina, 105

Country Index

Australia, 83
Austria, 106, 117

Belgium, 65, 120
Bulgaria, 24, 75

China, 24

Denmark, 9, 11, 20, 30, 40, 77, 81, 103, 108, 116, 127, 132

Estonia, 38, 51, 119

Finland, 28, 46, 59, 71, 87, 91, 101, 106, 110, 117, 118, 127

Germany, 57, 69, 79, 115
Great Britain, 6, 125

Israel, 79

Japan, 44, 104

Latvia, 44, 53, 85, 89, 105

Netherlands, 18, 40, 61, 106, 116, 117
Norway, xvi, 3, 9, 13, 24, 34, 65, 73, 75, 107, 108, 112, 113, 133, 134

Poland, 36

Sweden, 15, 26, 32, 42, 48, 55, 59, 63, 67, 73, 93, 95, 97, 107, 120, 121, 127
Switzerland, 22, 81, 130

United States of America, 4, 26, 53, 95, 125

Institution Index

Aalto University, 59, 106, 110, 117, 118
Aarhus University, 11, 30, 40, 77, 108, 116, 127
Alan Turing Institute London, The, 6
Austrian Academy of Sciences, 106, 117

Ben-Gurion University of the Negev, 79
Berlin-Brandenburg Academy of Sciences and Humanities, 115
Bern University of Applied Sciences, 22

Chalmers University of Technology, 107

Danish National Archives, 9

Edith Cowan University, 83

Finnish National Gallery, 59

Ghent University, 120
Gothenburg University, 67
Gävle University, 55

Indiana University Bloomington, 26
International Internet Preservation Consortium, 125

KU Leuven, 65

Linnaeus University, 127

National Institute of Advanced Industrial Science and Technology, 44
National Library of Estonia, 38, 119
National Library of Latvia, 85
National Library of Norway, 113, 133
National Museum of Japanese History, 104
Nicolaus Copernicus University, 36
NORCE Norwegian Research Centre, 13

Peking University, 24

Royal Netherlands Academy of Arts and Sciences, The, 18, 61

Society for Danish Language and Literature, 30
Southern Cross University, 83
Stanford University, 95
Swedish Institute for Language and Folklore, 121
Swedish National Heritage Board, 73
Swiss National Data and Service Center for the Humanities, 22

Takin.solutions, 24, 75
Technical University of Darmstadt, 79
Technical University of Denmark, 81
The British Library, 125
The Finnish Literature Society, 59
The Library of Congress, 125
The Royal Danish Library, 132

UiT The Arctic University of Norway, 9
Umeå University, 63
University of Agder, 107
University of Amsterdam, 18, 40
University of Basel, 22, 130
University of Bergen, 3, 13, 24, 34, 75, 134
University of Bern, 22
University of Borås, 42
University of California, 53
University of Copenhagen, 9, 20, 103
University of Gothenburg, 15, 26, 32, 42, 93
University of Helsinki, 28, 59, 71, 87, 106, 110, 117, 118, 127
University of Jyvaskyla, 91
University of Latvia, 44, 53, 89
University of Notre Dame, 83
University of Oslo, 65, 73, 108
University of Regensburg, 57, 69
University of Tartu, 51, 119
University of Turku, 46, 101
University of Utah, 4
University of Zurich, 81
Uppsala University, 48, 55, 59, 63, 95, 120, 121, 127
Utrecht University, 18, 116

Ventspils University of Applied Sciences, 105

Vrije Universiteit Amsterdam, 106, 117

Western Norway University of Applied Sciences, 112
Wikimedia Sverige, 97

Østfold University College, 13