Machine Learning and Data Mining
for
Digital Scholarly Editions

University of Rostock, 9-10 June, 2022

Conference Reader

Contact

For questions about the program, registration, travel, venue, etc., you can contact the organizing team at the following email address: ml-dse@i-d-e.de

Website

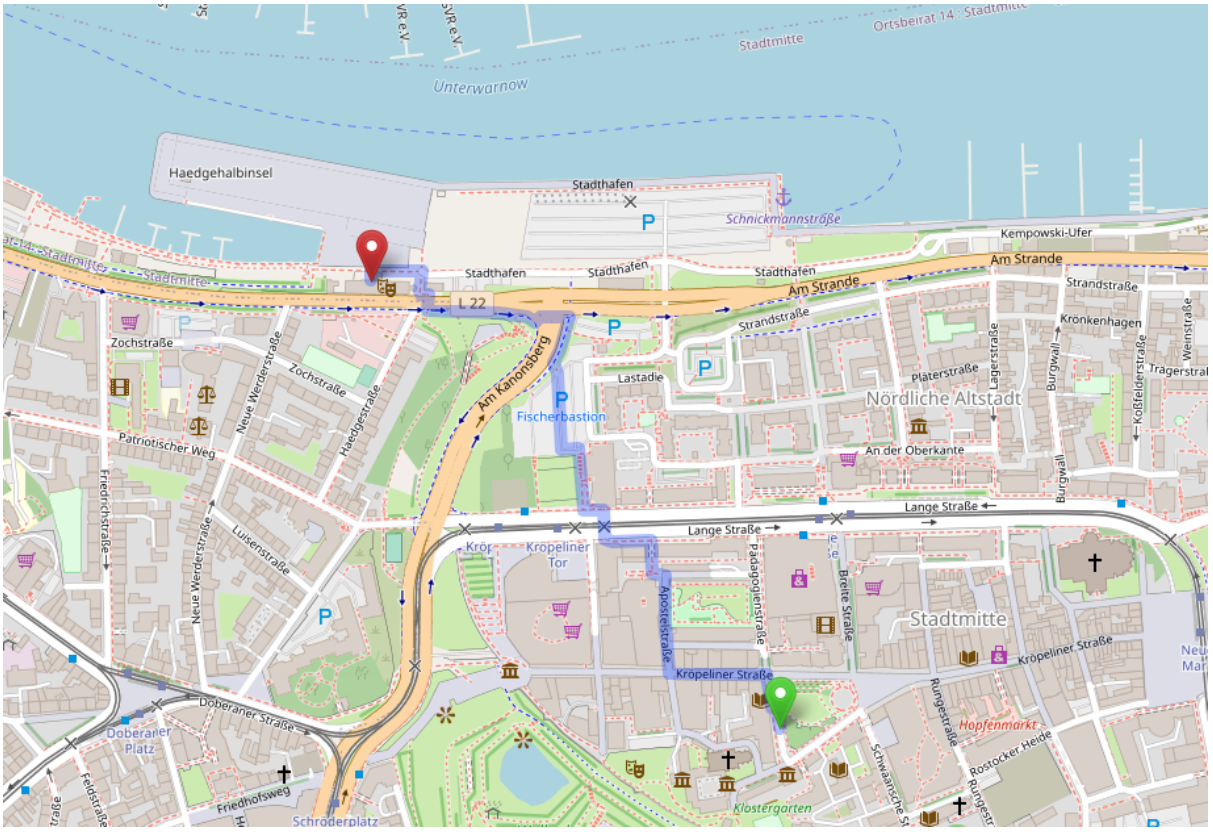
Information about the conference is also available at https://www.i-d-e.de/ml-dse

Venues

The conference will be held at the University of Rostock, in Northern Germany. The talks take place in room HS 218 in the main university building and the coffee breaks in room SR 217 in the same building (Universitätshauptgebäude, Universitätsplatz 1, 18055 Rostock). For dinner together on Thursday evening we will meet at the brewery and restaurant *Zum Alten Fritz* (Warnowufer 65, 18057 Rostock, https://www.alter-fritz.de/). See also the maps below.

Universitätshauptgebäude / Main university building

Route from the main station (Rostock Hauptbahnhof) to the main university building (Universitätshauptgebäude), about 20 minutes walking



Walk from the main university building to the restaurant *Zum Alten Fritz* (900 m, 15 minutes walking)

Program

Thursday, 9 June	
1.30 p.m.	Arrival
1.50 p.m.	Welcoming
2-3.30 p.m.	Moderation: Martina Scholger Tobias Hodel (University of Bern) Machine Learning Approaches to Making and Using Digital Editions: Premodern Documents, Text, and Named Entity Recognition Richard Hadden, Nina Rastinger, Matthias Schlögl (Austrian Academy of Sciences) Data Mining and Data Sowing: Automated Methods for Building a Digital Scholarly Edition of Historical Newspapers
3.30-4 p.m.	Coffee break
4-5.30 p.m.	Moderation: Ulrike Henny-Krahmer Sandra Bläß, Marie Flüh, Julia Nantke (Hamburg University), David Maus (University Library Hamburg) Quality Management for Machine Generated Data in Digital Scholarly Editions – Possibilities and Challenges Daniel Stökl Ben Ezra (École Pratique des Hautes Études), Hayim Lapin (University of Maryland) From HTR to Digital Critical Scholarly Edition: Reflexions on the Use of Machine Learning, Computational and Digital Humanities in the Sofer Mahir Project
5.30-6 p.m.	Break
6 p.m.	Keynote: Roger Labahn (University of Rostock) Machine Learning & Digital Humanities – a personal perspective
8 p.m.	Dinner together (Zum Alten Fritz)

Friday, 10 June	
9-10.30 a.m.	Moderation: Marc Lemke Werner Scheltjens (University of Bamberg) Extracting Preindustrial Logistics Patterns From a Digital Edition: Reflections on a Workflow for Data Mining the Schenkenschans Customs Registers (1630-1810) Manuel Portela (University of Coimbra) Text Mining the Book of Disquiet for Critical and Creative Explorations
10.30-11 a.m.	Coffee break
11 a.m.-12.30 p.m.	Moderation: Martina Scholger Ulrike Henny-Krahmer (University of Rostock), Frederike Neuber (Berlin-Brandenburg Academy of Sciences and Humanities) Topic Modeling in Digital Scholarly Editions Marc Lemke, Konrad Sperfeld, Jochen Zöllner (University of Rostock) Introducing NTEE: An Easy to Use Tool to Enrich TEI Files With Entities Based on State of the Art Neural Networks
12.30-13 p.m.	Closing discussion
13 p.m.	Lunch together / Goodbye

Abstracts

Sandra Bläß, Marie Flüh, Julia Nantke (Hamburg University), David Maus (University Library Hamburg)

Quality Management for Machine Generated Data in Digital Scholarly Editions – Possibilities and Challenges

In our paper, we present the automatisation-oriented workflow we apply in our project *Dehmel digital* to transcribe and index 35,000 documents (mainly handwritten letters). Our approach includes image digitisation, data transfer, HTR and transcription, tokenizing, NER, and disambiguating the entities. We combine manual, semi-automatic, and automatic resp. algorithm-driven methods by using diverse Machine Learning techniques. The goal is to process the enormous amount of analog letters so they can be presented in a DSE. A main issue throughout the whole process is the question of how we assure scholar quality of all of our produced data. By applying the quality management dimensions of Askham et al. – Completeness, Uniqueness, Timeliness, Validity, Accuracy, Consistency –, we examine in our paper how to measure and define quality at each work step, at which points in the workflow corrective interventions are possible without rendering the data structure unusable for subsequent processing steps, and how to deal with inaccuracies. A key interest here is also the role of the editor: How does it change when editorial work becomes algorithm-driven?

*****

Richard Hadden, Nina Rastinger, Matthias Schlögl (Austrian Academy of Sciences)

Data Mining and Data Sowing: Automated Methods for Building a Digital Scholarly Edition of Historical Newspapers

This article presents ongoing work on the DIGITARIUM, a project to digitise the *Wien[n]erisches Diarium* newspaper. The information contained within the *Diarium* is an invaluable resource for historical researchers, as well as for linguistic and other analysis. Here, we consider the process of digitisation, and the progress towards annotating the digitised texts, both using machine learning technologies. While this article focuses on the digitised texts themselves, considering them as a form of digital scholarly edition, both the process of creation and augmentation, and digital humanities-based research into the text can be seen as interdependent (even symbiotic): historical research employing machine learning techniques is fed back into the edition itself, forming layers of annotation, while simultaneously improving machine learning models. Such an approach raises questions regarding traditional approaches to textual scholarship and the status and utility of such a digital edition, which we shall consider.

*****

Ulrike Henny-Krahmer (University of Rostock), Frederike Neuber (Berlin-Brandenburg Academy of Sciences and Humanities)

Topic Modeling in Digital Scholarly Editions

Topic Modeling is a quantitative method of text analysis that has been increasingly used in recent years for the analysis of digital text collections in the humanities. In this paper, we investigate to what extent this also applies to text corpora that are being created as scholarly digital editions. Possible fields of application of topic modeling for the description and representation of contents in digital editions are discussed and from the perspective of the topic modeling workflow it is worked out which aspects have to be considered in particular when applying the method to scholarly edited texts. The method is applied to two corpora of correspondence of the German-language authors Jean Paul (1763-1825) and Uwe Johnson (1934-1984), which are being compiled in edition projects at the Berlin-Brandenburg Academy of Sciences and Humanities (BBAW). Topic modeling stands for a way of working with text that in many respects contrasts with critical editing. Nevertheless, the method can enrich the work on and presentation of digital scholarly editions in a thematic dimension if the different approaches to texts are explained and taken into account.

*****

Tobias Hodel (University of Bern)

Machine Learning Approaches to Making and Using Digital Editions: Premodern Documents, Text, and Named Entity Recognition

Machine learning capabilities and scholarly edition workflows seem to fit almost perfectly together at first sight. Text and entities (like persons and places) must be recognized and digitally processed. The presented paper focuses on premodern documents and depicts the current state-of-the-art in combining machine learning approaches for handwritten text and named entity recognition. Both text recognition and language model-based entity recognition are still developmental or need specific training when applied to unformalized (read *premodern*) textual entities; statistical scores are relatively low compared to standardized systems for modern languages. Ongoing developments in the digital humanities and the understanding to use available technology cautiously make for an optimistic outlook, even for languages and language forms with only sparse data reservoirs.

The use of deep learning offers in parallel insights into the consequences for scholarly editions, whereby the need to focus on deliberately chosen documents is minimized, and post-processing steps require to be intensified. The use of machine learning for editions leads thus to a reflective stance about digital scholarly editions and the insight that cooperation is critical in developing robust and consciously built machine learning systems that benefit the academic edition community.

*****

Marc Lemke, Konrad Sperfeld, Jochen Zöllner (University of Rostock)

Introducing NTEE: An Easy to Use Tool to Enrich TEI Files With Entities Based on State of the Art Neural Networks

In this paper we introduce the open-source software NTEE which is designed for automatically enriching XML-TEI files with XML elements as part of a Named Entity Recognition (NER) and Named Entity Linking (NEL) task within a Digital Scholarly Edition workflow. As our starting point for the development of the software, we first outline the fact that NER and NEL are desired supporting technologies in digital edition projects, but are rarely used due to a lack of expertise and the lack of adaptability in existing software solutions to edition-specific guidelines. Referring to this, we then show how NTEE is specifically designed for adaptability and accessibility: Based on state of the art neural language models, users can train and then apply their own NER taggers to XML-TEI files. For the subsequent manual postprocessing, NTEE offers technical support on the one hand for evaluating the achieved annotation results and on the other hand by accessing databases using Semantic Web technology in order to clearly identify the recognized entities. All this together can be handled in NTEE by the users within a single graphical user interface.

*****

Manuel Portela (University of Coimbra)

Text Mining the Book of Disquiet for Critical and Creative Explorations

The *LdoD Archive* is a highly dynamic archive dedicated to Fernando Pessoa's *Book of Disquiet*. It includes a scholarly edition of the work based on the autograph witnesses, a transcription of four canonical editions, and several interactive functions that allow subjects to explore multiple reading sequences, compare editorial versions, produce virtual editions of the work, and annotate texts in those virtual editions. These subject-oriented interactions are based on a combination of human-only and machine-assisted operations. This article discusses those machine-assisted interventions as forms of text mining for critical and creative explorations of the processes of reading and editing as interpretative actions. Thus the focus of the *LdoD Archive* is not on machine learning as a form of automation of text analysis *per se*, but rather on the uses of algorithmic procedures as contributions to the role-playing and gamification rationale of our editorial model.

*****

Werner Scheltjens (University of Bamberg)

Extracting Preindustrial Logistics Patterns From a Digital Edition: Reflections on a Workflow for Data Mining the Schenkenschans Customs Registers (1630-1810)

This paper discusses preliminary results and forthcoming tasks of a project that started in the fall of 2021 at the University of Bamberg. The project aims to facilitate the study of logistics patterns in German-Dutch transport and trade on the River Rhine in the early modern period by means of a digital edition of the customs registers of the Schenkenschans (SSZ). Based on an ongoing pilot study with a sample of the SSZ registers, the paper discusses the use of tools for HTR as starting point for creating a digital edition of non-narrative sources such as the SSZ customs registers and discusses the possibility of using machine learning for the standoff annotation of register entries. The pilot study aims to find out whether a digital edition is at all feasible for the SSZ registers and what role automated procedures may play during the process.

*****

Daniel Stökl Ben Ezra (École Pratique des Hautes Études), Hayim Lapin (University of Maryland)

From HTR to Digital Critical Scholarly Edition: Reflexions on the Use of Machine Learning, Computational and Digital Humanities in the Sofer Mahir Project

The present paper will give an overview about the use of machine learning as well as vintage algorithms from natural language processing (NLP) and computer vision (CV) and diverse digital humanities (DH) tools in the preparation of a series of digital editions of the major texts from the earliest phase of rabbinic Judaism. We employed these methods for page layout segmentation (PLS), handwritten text recognition (HTR), tokenization, lemmatization, named entity recognition (NER), text-reuse detection and text alignment. We are currently in the final stages of the data curation and preparation of the TEI-Publisher customization for the visualisation of the data.

*****

Bios (presenters and organizers)

Sandra Bläß

Since 2020 research assistant in the project *Dehmel digital*, University of Hamburg
2017–2019 student assistant, 2019–2020 research assistant in the project *forTEXT*, University of Hamburg
Areas of interest: narrative german literature of the 20th/21st century, digital literary studies, narrative concepts of identity, cultural studies of literature

Marie Flüh

Research assistant at the Institute for German Studies, University of Hamburg.
Since 2020 in the project *Dehmel digital*, 2018–2020 in the project *forTEXT*.
Areas of Interests: Computational Literary Studies, Gender Studies, methods of Digital Humanities and their transfer into teaching, emotions in literary texts.

Bernhard Geiger

studied electrical engineering at Graz University of Technology, Austria. He was a Senior Scientist and Erwin Schrödinger Fellow at the Institute for Communications Engineering, Technical University of Munich, Germany (2014 to 2017) and at the Signal Processing and Speech Communication Laboratory, Graz University of Technology, Austria (2017 to 2018). Bernhard is currently a Senior Researcher at Know-Center GmbH, Graz, Austria, where he also leads the Machine Learning Group. His research interests cover information theory for machine learning, theory-assisted machine learning, and information-theoretic model reduction for Markov chains and hidden Markov models.

Ulrike Henny-Krahmer

Ulrike Henny-Krahmer studied Regional Sciences of Latin America at the Universities of Cologne and Lisbon. Since October 2021, she is a junior professor for Digital Humanities at the University of Rostock. She is one of the managing editors of the journal “RIDE – A review journal for digital editions and resources”, published by the IDE. Her research focuses on digital scholarly editing, text analysis, and evaluation and sustainability of digital research output.

Tobias Hodel

Tobias Hodel is Assistant Professor in Digital Humanities at the University of Bern, Switzerland. His research interests include the theory of the digital humanities, machine learning in the humanities and critical algorithm studies.

Fabian Kaßner

Fabian Kaßner is a research associate in the academy project *Uwe Johnson-Werkausgabe* at the University of Rostock, where he is responsible for the technical infrastructure, data modeling (TEI) and web development of the digital edition Uwe Johnson: digitale Werkausgabe. He has a background in History, German Studies and Editorial Studies.

Roger Labahn

Roger Labahn studied mathematics at the University of Rostock, where he completed his doctorate in 1987 and his habilitation in 1994 in the field of Discrete Mathematics.
His move to the area of mathematical optimization at the end of the nineties was accompanied by an increased orientation towards application-oriented topics. In connection with long-term cooperations with PLANET GmbH, this increasingly focused on machine learning at a relatively early stage.
Since the end of the nineties, this work could be established in various projects, which eventually led from the state level to two Horizon2020 projects on the European level. This work has been accomplished in his project group CITlab (Computational Intelligence Technology Lab).
Here, the research focuses on the mathematical foundations of Machine Learning, while the applications are primarily concerned with the analysis of text documents or images in general. Starting with the automated recognition of layout and text, the spectrum has recently expanded towards content interpretation and Natural Language Processing in particular.
Since 2016, Roger Labahn has been an adjunct professor of mathematics at the University of Rostock.

Hayim Lapin

Hayim Lapin (Phd, Columbia University, 1995) is Robert H. Smith Professor of Jewish Studies and Professor of History at the University of Maryland, and former Director (Chair) of the Meyerhoff Center and Program for Jewish Studies. His research focuses on ancient history, and in addition is the co-director (with Stökl) of the e-rabbinica project, which to date

has produced multi-witness digital edition of the Mishnah and an alignment between the Mishnah and the Tosefta.

Marc Lemke

Marc Lemke, M.A. studied German literature and linguistics as well as modern and recent history at the University of Rostock. Since 2014 he has been involved in the work on the Digital Scholarly Edition of the works, letters and writings of Uwe Johnson at the University of Rostock. In this context, he is engaged in the development of methods for the computer-aided analysis of historical documents as part of the EU project NEISS (2019-2022).

David Maus

Since 2019 Head of Research & Development at State- and University Library Hamburg
IT architect of the project *Dehmel digital* (2020–2023)
2010–2019 Staff member at the Herzog August Bibliothek Wolfenbüttel, contributing to various DH and digitization projects
Areas of interest: Markup technologies, information architecture, hypertext/hypermedia, linked open data

Julia Nantke

Since 2019 Juniorprofessor for Modern German Literature with a focus on Digital Humanities and written artifacts at Hamburg University
PI of the project *Dehmel digital* (2020–2023)
2016–2019 Postdoctoral researcher at the DFG research training group 2196 “Document – Text – Editing”, University of Wuppertal
2016 PHd
Areas of research: Digital Literary Studies, materiality and mediality of literature, scholarly editing, literary theory, digital literature, literature and art of the Avantgardes

Frederike Neuber

Frederike Neuber is a textual scholar and digital humanist, working at the Berlin-Brandenburg Academy of Sciences and Humanities where she coordinates several digital edition projects. She is co-editor of the digital edition of the correspondence of the German writer Jean Paul (1763–1825), member of the Institute of Documentology on Scholarly Editing and one of the managing editors of the journal “RIDE – A review journal for

digital editions and resources". Her research interests include digital editions, (digital) text theories and evaluation practices for digital research.

Manuel Portela

Manuel Portela is Professor of English and Director of the PhD Programme in Materialities of Literature at the University of Coimbra. His research addresses writing and reading media and how they impact on literary forms and practices. The most significant results of his work can be seen in *Scripting Reading Motions: The Codex and the Computer as Self-Reflexive Machines* (MIT Press, 2013), *LdoD Archive: Collaborative Digital Archive of the Book of Disquiet* (2017-2022, https://ldod.uc.pt), co-edited by António Rito Silva, and *Literary Simulation and the Digital Humanities: Reading, Editing, Writing* (Bloomsbury, 2022).

Werner Scheltjens

Werner Scheltjens (1978) obtained his PhD in history at the University of Groningen (2009) and his Habilitation (with venia for social and economic history and East European history) at the University of Leipzig in 2020. Since January 2021, he is professor of digital history at the University of Bamberg, Germany. His research focuses on the application of digital historical methods to conduct research on preindustrial economic history, maritime history, and historical metrology.

Gerlinde Schneider

Gerlinde Schneider is a freelance software developer and digital humanist focussing on language technologies, digital scholarly editions and web engineering. Since 2020 she has been a member of the Institute for Documentology and Scholarly Editing (IDE). From 2012 to 2021 she was a developer and project coordinator at the Centre for Information Modelling at the University of Graz, where she contributed to several research and infrastructure projects. Currently, she is pursuing a qualification in Informatics in Graz.

Martina Scholger

is senior scientist at the Centre for Information Modelling – Austrian Centre for Digital Humanities at the University of Graz. She studied art history and completed her PhD in Digital Humanities in 2018. She has been a member of the Institute for Documentology and Scholarly Editing (IDE) since 2014 and a member of the TEI Technical Council, where she is currently serving as Chair, since 2016. Her research focuses on digital scholarly editing, semantic web technologies, and text analysis.

Konrad Sperfeld

Dr rer. nat. Konrad Sperfeld is a researcher in the field AI, specialized to Natural Language Processing. He works in the CITlab-Group of the Institute of Mathematics in the University of Rostock, which is specialized to develop Machine Learning technology. He is currently leading a group of young scientists in the NEISS Project, which is working on the partial automation of the creation of digital editions. He has 3 children, is married and lives in Rostock.

Daniel Stökl Ben Ezra

Daniel Stökl Ben Ezra (PhD, Jerusalem, 2002) is Chair of *Hebrew and Aramaic language, literature, epigraphy and paleography chair (4th century BCE - 4th century CE)* at the EPHE, PSL and member of the AOrOc laboratory (UMR 8546, PSL-CNRS). His research focuses on the Dead Sea Scrolls, ancient rabbinical literature, as well as digital and computational humanities. He is co-director (with P. Stokes) of the eScriptorium infrastructure for computational analysis of handwritten objects and co-director (with Lapin) on the e-rabbinica project.

Jochen Zöllner

Jochen Zöllner, M.S. studied physics and electrical engineering at the University of Rostock and works on machine learning based document processing technologies since 2017.
He is working at Planet AI GmbH Rostock and the NEISS Project as an AI Researcher writing his PhD thesis.

Scientific Committee & Organization

Scientific Committee

Helena Bermúdez-Sabel (University of Neuchâtel)
Hannah Busch (Royal Netherlands Academy of Arts & Sciences (KNAW))
Katrin Dennerlein (University of Würzburg)
Bernhard Geiger (Know-Center Graz)
Denis Helic (Graz University of Technology)
Ulrike Henny-Krahmer (University of Rostock)
Tobias Hodel (University of Bern)
Fabian Kaßner (University of Rostock)
Roman Kern (Know-Center Graz)
Mike Kestemont (University of Antwerp)
Marc Lemke (University of Rostock)
Fotis Jannidis (University of Würzburg)
Manuel Portela (University of Coimbra)
Patrick Sahle (University of Wuppertal)
Gerlinde Schneider (University of Graz)
Martina Scholger (University of Graz)
Georg Vogeler (University of Graz)

Organizers

Bernhard Geiger (Know-Center Graz)
Ulrike Henny-Krahmer (University of Rostock)
Fabian Kaßner (University of Rostock)
Marc Lemke (University of Rostock)
Gerlinde Schneider (University of Graz)
Martina Scholger (University of Graz)

The conference is co-organized by the Institut für Dokumentologie und Editorik, the Academy Junior Professorship for Digital Humanities at the University of Rostock, the Know-Center GmbH and the Centre for Information Modelling at the University of Graz. It is funded by the University of Rostock and supported by the NEISS project.