university of groningen

University Medical Center Groningen

University of Groningen

Computer Vision and Architectural History at Eye Level

Mager, Tino; Khademi, Seyran; Siebes, Ronald; van Gemert, Jan; De Boer, Victor; Löffler, Beate; Hein, Carola

Published in:
Mixing Methods

IMPORTANT NOTE: You are advised to consult the publisher's version (publisher's PDF) if you wish to cite from it. Please check the document version below.

Document Version
Publisher's PDF, also known as Version of record

Publication date:
2023

Link to publication in University of Groningen/UMCG research database

Citation for published version (APA):
Mager, T., Khademi, S., Siebes, R., van Gemert, J., De Boer, V., Löffler, B., & Hein, C. (2023). Computer Vision and Architectural History at Eye Level: Mixed Methods for Linking Research in the Humanities and in Information Technology. In B. Schneider, B. Löffler, T. Mager, & C. Hein (Eds.), *Mixing Methods: Practical Insights from the Humanities in the Digital Age* (pp. 123-144). (Digital Humanities Research; Vol. 7). Bielefeld University.

Copyright
Other than for strictly personal use, it is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license (like Creative Commons).

The publication may also be distributed here under the terms of Article 25fa of the Dutch Copyright Act, indicated by the “Taverne” license. More information can be found on the University of Groningen website: https://www.rug.nl/library/open-access/self-archiving-pure/taverne-amendment.

Take-down policy
If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Downloaded from the University of Groningen/UMCG research database (Pure): http://www.rug.nl/research/portal. For technical reasons the number of authors shown on this cover page is limited to 10 maximum.

Download date: 01-11-2023

BIRGIT SCHNEIDER, BEATE LÖFFLER,
TINO MAGER, CAROLA HEIN (EDS.)

MIXING METHODS

PRACTICAL INSIGHTS FROM THE HUMANITIES IN THE DIGITAL AGE

DIGITAL HUMANITIES RESEARCH
BIELEFELD UNIVERSITY PRESS

Birgit Schneider, Beate Löffler, Tino Mager, Carola Hein (eds.)
Mixing Methods

Digital Humanities Research | Volume 7

Editorial

Digital Humanities is an evolving, cross cutting field within the humanities employing computer based methods. Research in this field, therefore, is an interdisciplinary endeavor that often involves researchers from the humanities as well as from computer science. This collaboration influences the methods applied as well as the theories underlying and informing research within those different fields. These implications need to be addressed according to the traditions of different humanities' disciplines. Therefore, the edition addresses all humanities disciplines in which digital methods are employed. **Digital Humanities Research** furthers publications from all those disciplines addressing the methodological and theoretical implications of the application of digital research in the humanities.
The series is edited by Silke Schwandt, Anne Baillot, Andreas Fickers, Tobias Hodel and Peter Stadler.

Birgit Schneider is a professor for knowledge cultures and media environments at the Department of European Media Studies at Universität Potsdam, Germany. She studied art and media studies as well as media art and philosophy in Karlsruhe, London and Berlin.
Beate Löffler researches and teaches at Technische Universität Dortmund. She received an engineering degree in architecture in Potsdam and studied history and art history in Dresden afterwards.
Tino Mager is an assistant professor of the history and theory of architecture and urbanism at the University of Groningen. He studied media technology in Leipzig and art history and communication science in Berlin, Barcelona and Tokyo.
Carola Hein is a professor of the history of architecture and urban planning at Delft University of Technology. She studied architecture and urban planning in Hamburg and Brussels.

Birgit Schneider, Beate Löffler, Tino Mager, Carola Hein (eds.)

Mixing Methods

Practical Insights from the Humanities
in the Digital Age

[transcript]

Bibliographic information published by the Deutsche Nationalbibliothek
The Deutsche Nationalbibliothek lists this publication in the Deutsche Nationalbibliografie; detailed bibliographic data are available in the Internet at http://dnb.d-nb.de

This work is licensed under the Creative Commons Attribution 4.0 (BY) license, which means that the text may be remixed, transformed and built upon and be copied and redistributed in any medium or format even commercially, provided credit is given to the author.
Creative Commons license terms for re-use do not apply to any content (such as graphs, figures, photos, excerpts, etc.) not original to the Open Access publication and further permission may be required from the rights holder. The obligation to research and clear permission lies solely with the party re-using the material.

First published in 2023 by Bielefeld University Press, Bielefeld
An Imprint of transcript Verlag https://www.transcript-verlag.de/bielefeld-up
© Birgit Schneider, Beate Löffler, Tino Mager, Carola Hein (eds.)

Cover layout: Maria Arndt, Bielefeld
Printed by: Majuskel Medienproduktion GmbH, Wetzlar
https://doi.org/10.14361/9783839469132
Print-ISBN: 978-3-8376-6913-8
PDF-ISBN: 978-3-8394-6913-2
ISSN of series: 2747-5476
eISSN of series: 2749-1986

Printed on permanent acid-free text paper.

Contents

Preface 9

I MIXED METHODS

Mixing Methods. Practical Insights from the Humanities in the Digital Age
Birgit Schneider, Beate Löffler, Tino Mager, Carola Hein 13

Mixed Methods and the Digital Humanities
Andrew Prescott 27

II TEN CASE STUDIES

#PARAPHRASE 45

Innovation in Loops: Developing Tools and Redefining Theories within the Project 'Digital Plato' (Digital Plato)
Eva Wöckener-Gade, Marcus Pöckelmann 47

#SIMILARITY 61

Reading at Scale. A Digital Analysis of German Novellas from the 19th Century (Reading at Scale)
Thomas Weitin, Simon Päpcke, Katharina Herget, Anastasia Glawion, Ulrik Brandes 63

#CORPUS 79

On Designing Collaboration in a Mixed-Methods Scenario. Reflecting Quantitative Drama Analytics (QuaDrama)
Janis Pagel, Benjamin Krautter, Melanie Andresen, Marcus Willand, Nils Reiter 81

#HUMAN-IN-THE-LOOP 103

Dhimmis and Muslims – Analyzing Multi-Religious Spaces in the Medieval Muslim World (DhiMu)
Ralph Barczok, Max Franke, Steffen Koch, Dorothea Weltecke 105

#VISUALIZATION 123

Computer Vision and Architectural History at Eye Level: Mixed Methods for Linking Research in the Humanities and in Information Technology (ArchiMediaL)
Tino Mager, Seyran Khademi, Ronald Siebes, Jan van Gemert, Victor de Boer, Beate Löffler, Carola Hein 125

#CANON 145

Musical Schemata: Modelling Challenges and Pattern Finding (BachBeatles)
Markus Neuwirth, Christoph Finkensiep, and Martin Rohrmeier 147

#MODELLING 165

Free Verse Prosodies: Identifying and Classifying Spoken Poetry Using Literary and Computational Perspectives (Rhythmicalizer)
Timo Baumann, Hussein Hussein, Burkhard Meyer-Sickendiek 167

#MACHINE LEARNING 187

Interpreting Climate Images on the Internet: Mixing Algorithmic and Interpretive Views to Enable an Intercultural Comparison (ANCI)
Birgit Schneider, Thomas Nocke, Paul Heinicker, Janna Kienbaum 189

#QUANTIFICATION 213

Detecting Authorship, Hands, and Corrections in Historical Manuscripts. A Mixed-methods Approach towards the Unpublished Writings of an 18th Century Czech Emigré Community in Berlin (Handwriting)
Roland Meyer, Aleksej Tikhonov, Robert Hammel 215

#UNCERTAINTY 237

Encoding, Processing and Interpreting Vagueness and Uncertainty in Historical Texts – A Pilot Study Based on Multilingual 18th Century Texts of Dimitrie Cantemir (HerCoRe)
Cristina Vertan 239

#HETEROGENEITY 259

Authors 261

Preface

In November 2015, the Volkswagen Foundation put out a call, "'Mixed Methods' in the Humanities? – Support for Projects Combining and Synergizing Qualitative-Hermeneutical and Digital Approaches". The title of the call was quite cumbersome, describing the objective of the call. On the other hand, the title adopted the term 'mixed methods' from the social sciences to use it in the context of the humanities – a novelty at that time.

That is exactly the overall aim of the Volkswagen Foundation: we want 'to swim ahead of the tide', as it is described in our new funding strategy. The paradigm shift of data-intensive research in the humanities was identified at a very early stage; innovative funding opportunities were designed. In 1999, the Volkswagen Foundation launched the programme 'Documentation of Endangered Languages'. About 100 endangered languages were documented in an electronic archive freely available on the internet (https://dobes.mpi.n). The DobeS archive is considered today as FAIR data 'avant la lettre'. In order to support the innovative power of data-intensive research in the humanities, the Volkswagen Foundation in 2012 supported the foundation meeting of the regional association DHd – Digital Humanities im deutschsprachigen Raum. Later, the Foundation organized the international conference '(Digital) Humanities Revisited – Challenges and Opportunities in the Digital Age', followed by a conference on data quality organized together with the German Rat for Informationsinfrastrukturen. Lastly, in November 2021, the Foundation decided to integrate its various efforts and set a conceptual framework with an overall Open Science Policy.

The present publication 'Mixing Methods. Practical Insights from the Humanities in the Digital Age' also wants to meet the challenge of 'swimming ahead of the tide'. The Mixed Methods call was open to all disciplines in the humanities. Its overall aim was to prevent a confrontation between those scholars in the humanities working with non-digital approaches (from hermeneutic to structuralist, post-structuralist and postcolonial approaches) and humanist scholars working with digital methods. Instead of confrontation, its overall aim was to focus on the opportunity for great discoveries and intellectual innovation while integrating the two approaches. The project teams were asked not only to work on a specific research ques-

tion within their discipline, but also to reflect on the interface between the two procedures—digital/non-digital—at a theoretical-methodological level. How can the two approaches be linked and synergized when a shared object of research is being investigated? How do digital methods change the humanities epistemically? What can we learn from living the 'two worlds' at the same time? Nine projects were funded and a tenth project was integrated into the programme's network as it had quite a similar topic. A kick off meeting was organized followed by a midterm workshop where social scientist Udo Kuckartz, whose manual 'Mixed Methods. Methodologie, Forschungsdesigns und Analyseverfahren' (2014) originally provided the idea for the whole programme, presented in detail how to deal with mixing methods in his disciplinary field. At this workshop, the 10 project teams decided on publishing a joint volume. In putting together the results of the individual projects and their reflection at the theoretical-methodological level, this volume aspires to draft a new understanding of the humanities in the 21st century: What will the separate disciplines in the humanities have in common when they use digital approaches?

This publication took some time because of restrictions put on normal life due to the corona pandemic. Now it is our pleasure to present it as an invitation for interdisciplinary discovery. We thank the project teams for their sound cooperation. We thank Andrew Prescott, digital humanist from the very start, for elaborating on the history of digital humanities. And above all, we thank the editorial board – Birgit Schneider, Beate Löffler, Tino Mager and Carola Hein – for their tenacity to compose and complete this publication.

Vera Szöllösi-Brenig
Volkswagen Foundation

I MIXED METHODS

Mixing Methods. Practical Insights from the Humanities in the Digital Age

Birgit Schneider, Beate Löffler, Tino Mager, Carola Hein

Digitality is a cause and a consequence of different data cultures. It applies to the 10 research projects that are included in this volume. They are rooted in various humanities disciplines such as art history, philosophy, musicology, religious studies, architectural history, media studies, and literature studies. As diverse as the disciplines are the objects and their formats, which are the subject of this book. The cultural data of the projects include recordings of music and spoken word, photographs and other types of images, handwriting, typoscripts and maps. The oldest material dates back to 500 BCE, followed by medieval times, the 18th and 19th centuries, early 20th century and the present. All projects share that they study their material with digital methods, although digitality comes into play at different moments and layers in each of the projects. Hardly readable manuscripts from the 18th century have to be treated with specialized OCR-methods while Plato's texts are already available in digital form, and therefore open up other affordances for analysis. Special analysis possibilities had to be developed for certain image sources. For all projects, however, it is equally true that only the digitization of the objects makes them accessible to the methods that are the subject of this book.

If digitized cultural objects enable new research approaches, the question arises as to what benefit is actually produced when these objects are available in digital form. The additional value lies not only in the accessibility of data, but also in the questions that the digitized material allows us to ask, or which old questions can be answered in new ways on the basis of the digitized material. This means, digital cultural materials analysed by digital methods change the epistemological approaches of the humanities by opening up to research cultures formerly not used in the humanities. Not only are scientific methods relevant to the humanities, but the humanities themselves and their way to address the world are also relevant to science.

Much ink (or pixel) has already been spent defining what digital humanities are and are not.[1] We do not add to this but refer to philosopher Sybille Krämer's four aspects of digital humanities practice here. She writes that digital humanities projects involve "(1) The dataization of research subjects; (2) the use of either 'data-based' or 'data-guided' algorithmic research techniques; (3) the visualization of the results of analysis in a form that can be received by humans; (4) the novelty value of the findings."[2] As such the editors of this volume understand digital humanities as a highly explorative field, which tentatively researches what we can learn from digitization beyond the analogue and digital sources collected in databases and library catalogues and beyond the research we are already conducting in the humanities. It probes if we can fill research gaps with these new approaches or ask questions we have not asked until now. It is not about the humanities converting their data into digital logic, nor is it about computer science aligning its practices entirely with the concerns of the humanities. Instead, it is about finding a common ground. At the same time digital humanities, as we understand them, are calling for a general increase in digital literacy as a cultural technique in the humanities while not declaring interpretations as absolute. Against this backdrop, the digital transformation that is mirrored by the digital humanities is not a threat to the traditional theories, methodologies and disciplinary identities of the humanities. Rather it is asking humanities and computer science to get involved in new fields at eye level. However, we must admit that in practice, the two approaches very often stay unrelated or sometimes are even juxtaposed. The projects in this book start at this point of rupture by also reflecting self-critically on the question: what happens if both approaches are combined in one and the same research project by using a 'mixed methods' approach?

This question was the starting point for 10 research projects funded by the Volkswagen Foundation in the period of 2015–2020 with the funding line *'Mixed Methods' in the Humanities? – Support for Projects Combining and Synergizing Qualitative-Hermeneutical and Digital Approaches*. The projects, covering diverse fields of the humanities and using multiple digital methodologies, were encouraged to carry

1 See e.g. Jason Heppler, "What is Digital Humanities," accessed January 23, 2023, https://whatisdigitalhumanities.com/., Melissa Terras, "Quantified Digital Humanities," UCL Centre for Digital Humanities, accessed January 23, 2023, https://www.ucl.ac.uk/infostudies/melissa-terras/DigitalHumanitiesInfographic.pdf., Anne Burdick, Johanna Drucker, Peter Lunenfeld, Todd Pressner and Jeffrey Schnapp, *Digital Humanities* (Cambridge: MIT Press, 2016)., David M. Berry and Anders Fagerjord, *Digital Hiumanities: Knowledge and Critique in a Digitale Age* (Cambridge: Polity, 20).

2 Sybille Krämer, "Der 'Stachel des Digitalen' – ein Anreiz zur Selbstreflexion in den Geisteswissenschaften? Ein philosophischer Kommentar zu den Digital Humanities in neun Thesen," position paper of the keynote speech at the annual conference DHD 2018 at Cologne University.

out not only their individual research but also exploration of the interface between the two epistemological approaches at a theoretical and methodological level. All projects at the heart of this book use mixed methods approaches to conduct research in their various fields. The results of the projects are at the core of this volume. The research projects developed and used a wide range of methods to explore their research questions and their specific cultural data corpus. The term digital humanities, therefore, implies an abundance of very heterogeneous digital methods such as text mining, visual/auditive pattern detection, network analysis, statistics, visualization tools—which are themselves objects of humanistic interpretation. All these different methods are presented in this book and reflected upon in practice. We see it as strength of this book that we can place quite different disciplines side-by-side in order to jointly ask and compare how digital methods confront humanities subjects with different challenges.

By doing so, the book provides insights into concepts on how to work together in such projects in an interdisciplinary way. It addresses some of the most virulent questions in the field by exploring the potentials that arise from combining humanities issues with digital methods in the form of hands-on reports, productive findings and reflections. It contains an analysis of new terms that are emerging in practice and from co-productive teams, dealing with corpus as well as cultural data, methodologies including the status of machine learning as well as visualizations and, last but not least, new forms of collaboration.

Mixing methods, mixing research paradigms

In the first volume of this book series Silke Schwandt refers to the distinction "between computing methods being used '*for* and *in* the humanities"[3], as brought up by Edward Vanhoutte. The prepositions 'in' and 'for' express profoundly different relationships between methods and research fields. This simple discrimination, therefore, contains the challenges in the field, as mixing methods leads to mixing research paradigms, nothing more and nothing less. Thus, the idea of mixing methods is also about abandoning the strict and simplified dualism of qualitative and quantitative research in a productive way. The term mixed methods is relatively new. It was adopted from the social sciences during the last two decades to describe the intersection of the approaches. This type of research design was described and philosoph-

3 Silke Schwandt, *Digital Methods in the Humanities. Challenges, Ideas, Perspectives* (Bielefeld: Bielefeld University Press, 2020), 7.
Edward Vanhoutte, "The Gates of Hell: History and Definition of Digital | Humanities | Computing" in: Melissa M. Terras, Julianne Nvhan and Edward Vanhouette (eds.): *Defining Digital Humanities: A Reader* (London: Routledge, 2016), 120.

ically reflected upon in empirical social science most influentially by John Creswell. He uses the term to explain how to mix different methods "in all phases of the research process" and how in the process different "sets of beliefs" and "theoretical lenses", for example a post-positivist, a constructivist or a pragmatist worldview, lead to different methodologies.[4]

Originally developed for the analysis of social or natural science data, digital humanities projects can adapt the idea of designing a mixed method study that relates quantitative data to qualitative interpretations. It sounds like a "wild card" or even "epistemological anarchism" as reasoned by Paul Feyerabend in his essay *Against Method* as a means to scientific progress.[5] But it is not, because it opens up different methodologies, while the choice of methods is not arbitrary at all but cautiously and rigorously adjusted to the research questions and the research objects. The idea of mixing and composing methods in a research design has turned out to be very productive at different levels, although in another way than in the social sciences: Digital humanities do not correspond to quantitative social sciences; instead, the epistemological quality of digital humanities uses digital tools working on largely qualitative data. Cultural data, however, is not fact, but a cultural product and, moreover, a transformation of materiality. The need for new methods in the humanities is grounded in new research questions and the transformation and/or production of digital cultural data. The transformation of culture in data and networks at many levels, a process that has covered more and more areas during the last decades, can be answered in unprecedented ways with the help of digital methods. This means that a combination of traditional qualitative and innovative digital methods might address the problems and research questions concerning digital corpora in a novel way. It might even mean to deal with the societal questions of who owns culture and who monetises it, and what role open-access research can play.

Creswell combines different definitions of mixed methods by highlighting key characteristics of this research design such as the "researcher collects and analyses persuasively and rigorously both qualitative and quantitative data" or she/he "mixes (or integrates or links) the two forms of data concurrently by combining them (or merging them), sequentially by having one built on the other, or embedding one within the other."[6] The added value of this approach lies in the enhanced understanding it may give to a study, and also in the possibility that deficits of qualitative or quantitative methods can be balanced out by combining them. Creswell distinguishes various mixtures and sequences of research phases along a classification of

4 John Creswell, *Qualitative, quantitative and mixed methods approaches* (Thousand Oaks, California: SAGE, 2003), 208–225, 2 and 39.

5 Paul Feyerabend, *Against Method. Outline of an anarchist theory of knowledge* (London: Verso, 2010).

6 Creswell, *Qualitative, quantitative and mixed methods approaches*, 5.

their methods such as "sequential design", "multiphase design", "embedded design" or "exploratory design".[7] The types differ in how and when they collect data and mix methods. For example, a research project may start with qualitative research and then, in a second sequence, transform the research to quantitative methods, or the other way round. Both parts can play a balanced role or one part is given more weight than the other gets.[8] It becomes obvious that the idea and practice of mixed methods is challenging for all research paradigms.

The projects in this book took the idea of mixing methods as a starting point, and adapted the research designs to their needs. At the same time, all research projects productively question this demarcation and the clean separation between approaches and methods. The idea of mixed methods is the guiding principle of the research projects gathered here. They systematically combine qualitative and quantitative approaches, without immediately assigning one approach or paradigm to the humanities or computer science. Computer scientists have recourse to hermeneutic interpretation and humanities scholars also process data. This introduction presents the different approaches and critically reflects on the epistemic implications that emanate from them.

Cultural data and the culture of data

Humanities data are distilled from heterogeneous sources, specifically chosen and embedded in interpretative patterns that allow for a lot of greyscaling. While the words and ideas are precise, the connections between them are negotiable, evidently dependent on the question at hand and the perception of context. At the same time, digital data occupy the large space between the diversity of humanities and the often quantifiable and objectifiable dimensions of data in natural science and engineering. And yet, even here the epistemic and epistemological contexts are shaped by culture: the culture of native language (even if we speak English), the culture of teamwork, and the habits of the field, the department or the company. As such, digital humanities are less a meeting of different kinds of data or methods than one of different cultures of handling data and applying methods.[9]

We can problematize the very idea of data with Johanna Drucker, who wrote in the first volume of *DHQ* in 2011 that "*[c]apta* is 'taken' actively while *data* is assumed

7 Creswell, *Qualitative, quantitative and mixed methods approaches*, 208–225. See also Udo Kuckartz, *Mixed Methods. Methodologie, Forschungsdesigns und Analyseverfahren* (Wiesbaden: Springer, 2014), 81–83.

8 Kuckartz, *Mixed Methods*, 2014, summarizes these different types in Kuckart Chapter 2.

9 On this issue see Ludwik Fleck, *Genesis and Development of a Scientific Fact*, New Edition (Chicago, University of Chicago Press, 1981). Michel Foucault, "The Discourse of Language" trans. Rupert Swyer in: *Social Science Information* 10/2, 1971, 15–17.

to be a 'given' able to be recorded and observed. From this distinction, a world of differences arises. Humanistic inquiry acknowledges the situated, partial, and constitutive character of knowledge production, the recognition that knowledge is constructed, *taken*, not simply given as a natural representation of pre-existing fact."[10] Digital humanities projects are problematizing the fact that their data are constructed, because cultural data usually present objects which have not been distilled from technical instruments but have been produced by an author or artist in the first place. Therefore, cultural data never suit the idea of 'raw data' easily, although cultural data can be experimentally treated as raw data. This assessment of data as something unsharp and fluid saves digital humanities researchers from viewing their data through the positivist lens.

A classical differentiation between digital and analogue data is that while analogue data constitute a continuum, digital data present a sequence of states or signs. William J. Mitchell has given a comprehensible metaphor in 1992 already to mark the difference between the both: "The basic technical distinction between analogue (continuous) and digital (discrete) representation is crucial here. Rolling down a ramp is continuous motion, but walking down stairs is a sequence of discrete steps – so you can count the number of steps, but not the number of levels on the ramp."[11] This basic distinction can serve as a heuristic to differentiate the status of cultural objects that change not in their symbolic meaning but in their physical state through digitization, thus offering new starting points for research and reflection.

Digital humanities work with digital native or digitized material. Digital native data may be machine readable texts or digital photography. Digitized data may be digitized manuscripts, graphics, maps, analogue films or photographs or sound recordings. At a second level, digital objects are networked in digital archives—locally on a hard drive or globally in the internet through metadata. If cultural objects are converted into digital codes or if cultural objects are products of digital technology initially, new starting points for research open up because the very form of existence of digital objects offers a different interface for algorithmic and numeric operations. The starting point for the projects presented in this book is, therefore, a digital corpus of cultural material that has been assembled in many ways.

From reading the research chapters in his book, it becomes clear that the heuristic dualism of the digital and the analogue is not always satisfactory, nor is the dichotomy between quantitative and qualitative research. Digital, quantitative and non-digital, hermeneutic research methods are often seen as different perspectives. But, when addressing the mixed methods agenda, one of the most interesting findings has been that, in fact, they merge much more than usually thought of. Specif-

10 Johanna Drucker, "Humanities Approaches to Graphical Display" in: *DHQ*, 2011.

11 William Mitchell, *The Reconfigured Eye. Visual Truth in the Post-Photograph Era* (Cambridge, Mass. et al.: MIT Press, 1992), 4.

ically, it is hardly possible to define where traditional, hermeneutic research ends and where its digital counterpart begins. Both approaches comprise iterative processes, both are largely based on comparison and analogy, and both combine specific questions with overarching perceptions of relevance. However, in day-to-day work, epistemic and epistemological gaps become apparent as well and challenge the theoretical discussion when, for example, terms are used in very different ways and the expectations of what is productive outcome and insight differ profoundly. The project chapters and their case studies provide insight into the extent to which digital methods are simultaneously restrictive and productive in their programming – how they enable and limit cultural data analysis.

Interdisciplinary teams as a form of collaboration

Digital humanities projects are usually carried out in a team and are interdisciplinary by nature. They bring together researchers from backgrounds in the humanities and computer science. For this reason, there is the question of how the work of computer science and humanities is split and balanced in the projects and how these areas are mediated in the course of the research, without having to assign the role of an auxiliary science to one discipline. Interdisciplinarity is the core challenge of all complex research projects addressing contingent phenomena. It is essential to balance highly specialized expertise of the involved fields with the need for communication and coordination of research questions and results. In digital humanities, the parameters of interdisciplinarity between the humanities and computer science seem to be clearly delineated, with the latter often serving to solve the problems of the humanities while bringing the former to tears with the imperatives and prohibitions of ontological structures. Yet, the collaborations in the mixed methods projects show that we should think twice in this regard.

All the projects assembled in this volume set out to address digital humanities topics on a par with the resulting research relevant and pioneering for each of the involved fields. They set out to pool particular approaches and methods, ideas and questions and to build research practices from this pool. Their accounts provide insight into the additional workload originating in the need to agree on terms, meanings and the perception of relevance. They point to pitfalls and conceptual conflicts, such as in publication practices and in copyright uses, and to the establishment of new routines of validation or surprising insights into one's own subject. They also aim to engage with questions of societal justice, such as those related to the availability of specific types of data that have been conserved over centuries, while other types have disappeared. For example, information on colonial structures may be more prominent than about vernacular buildings, information on male actors is more easily available than on female. As such, the projects with their individual ways

of solving the task at hand prove to be laboratories of interdisciplinary convergence. Their experiences sketch ways to make such joint projects work epistemically and methodologically as well as spatially and organizationally.

A common denominator is the iterative processes of evaluation both in terms of questions, procedures and (preliminary) results. Alterations, often between human and technological contributions, and not necessarily only a human-in-the-loop set-up, ensure the ongoing validation of expectations, terminologies and outcomes. In some cases, this is achieved by setting up not only regular meetings to coordinate project parts but by actually co-working. Yet, the involved teams of most projects are inter-institutional and non-local: In some cases, academic and non-academic institutions cooperate. In others, the involved partners belong to different universities in different cities, even abroad. Against this background, setting up a joint space for cooperation, getting to know and growing to trust each other prove once more to be critical to the success of the proposed research. Thus, although digital humanities enable decentralized cooperation, they do not overcome the mechanisms of human collaboration rooted in actual social relations. It is frequently the trust in each other and an already well-established interdisciplinary cooperation in other contexts that allow stepping out of the established frameworks to address digital humanities and to mix methods.

This is not yet the end of the overarching findings from the comparison of the groups. In retrospect, many projects expressed their experiences of over- or underestimating challenges and/or possibilities concerning their partner disciplines. It referred not to the general feasibility but to issues of, for example, data quality and data complexity. Instead of hindering the work, it required a constant adjustment of resources towards the research goals. These experiences point once more to the epistemic dimensions of digital humanities as a still emerging field in which the negotiation of terms, concepts and aims is in a state of flux. Any shifts here influence the adjoining areas in computer science and humanities, ask for adaptations and mirror back resulting discourses. As such, the experienced interdisciplinarity circles the disciplinary habits far beyond digital humanities and mixed methods projects.

None of the projects failed.[12] Yet, the reports refer to critical points rooted in the very processes that enable digital humanities and the application of mixed methods. The critical one for exploratory research is the copyright for data. While the general use of digitized or digital data sets is often possible, the means needed to communicate the research itself and the sustainable availability of data and research for validation and continuation by other teams are frequently barred. The latter, however, is

12 On the issue of failures, see Dena Fam and Michael O'Rourke (ed.), *Interdisciplinary and Transdisciplinary Failures. Lessons Learned from Cautionary Tales* (Abingdon, Oxon; New York, NY: Routledge, 2020).

an indispensable part of quality management and routinely required by funding institutions. The resulting quandary forces research to go sideways to address relevant issues from a cumbersome angle or to ignore some topic entirely. Another observed issue, which has been overcome by all the projects in this funding line though, poses a challenge to the further development of explorative research in digital humanities, especially in the context of mixed methods. It is the hidden complexity. On the one hand, there is the impossibility of understanding the other discipline's cutting-edge research situation and, on the other, one's own ignorance concerning the intrinsic complexities of one's knowledge and order. For the implementation of mixed methods in digital humanities, it might mean to set up research proposals that are structurally different from the usual mono- or interdisciplinary projects, where the scope of conceptual overlapping is clearly known in advance.

Structure of the book

The book has a three-level structure. The first level of the book situates the digital humanities conceptually and historically and discusses the different approaches of mixed methods in the cooperation between humanities and computer science in general. Andrew Prescott provides a brief and concise history of the digital humanities. Beyond that, the focus is on general questions of methodology, on the debate as to how the involved fields change when they practise mixed methods, on what it means to work with uncommon data, and what it means when scholars present results obtained from data and research approaches uncommon in their own field.

The second level is a glossary of central terms for the research field of digital humanities. It discusses the following terms each with definitions provided by the *Oxford English Dictionary* and by the involved research projects, distilled from the essays themselves and from comments added during general discussions: *Paraphrase; Similarity; Corpus*; *Human-in-the-Loop-Approach; Visualization; Canons; Modelling; Machine Learning; Quantification*; *Uncertainty*; and *Heterogeneity*. The exchange is based on the experience of the involved research projects and points towards the complexity of concepts shaping digital humanities at the moment and creating subgroups as well as subtopics while cross-connecting in search of new approaches and fruitful cooperation. The terms are understood to provide bridges between the project chapters to demonstrate parallels and storylines. The glossary terms provide no final definitions but sketch a current state of discussion between the involved fields and foci. It is understood to underline the still ongoing negotiation of meaning for many of these terms that threaten to create new spheres of misunderstanding and disconnection.

The third level contains the 10 research chapters of the research projects involved. They range from literary studies to musicology, image studies, history of

religion, history of architecture, history of linguistics and text analysis. All chapters follow a similar scheme. They not only describe their approach, but also deal with their respective understanding of data or corpus, method or analysis. They describe in detail the team's collaboration, division of labour and processes. Here, we are especially interested in reflecting what is challenging in the projects, what the novel types of results are and what lead to disruptions in the process.

The work of the 10 projects assembled here includes different disciplinary spheres. The expertise from classical humanities such as literary studies and philology, regional linguistics and histories to visual, urban and media studies faces fields of competence in computer science concerned with machine learning and image recognition, corpus linguistics or interface design. Between them, some areas materialize that are already conceptually mixed and embody digital humanities, such as computational musicology and computational linguistics or information visualization, circumscribing the fluidity of the epistemic processes we observe today. In the sense of the funding line and this book, not only did the projects work towards the proposed results but they were also laboratories themselves. As such, the actual research results such as digital tools or conceptual insights are manifold and materialize largely in the relevant journals while the book compiles contributions that focus on the process-related outcomes of mixed methods. In this context, projects without overlapping content not infrequently share methodological challenges or digital humanities-related foci, as expressed in the glossary terms.

Digital Plato (Innovation in Loops: Developing Tools and Redefining Theories within the Project 'Digital Plato') focuses on the detection of paraphrases for Plato in ancient Greek sources. Since many of them neither referred to the source nor did they quote verbatim, the project came to reconsider concepts such as 'paraphrase' and 'intertextuality' in literature in its process to develop a tool for paraphrase search.

In contrast, the historical normativity of a literary corpus allowed the project *Reading at Scale* (A Digital Analysis of German Novellas from the 19th Century) to initiate an iterative process of operationalization towards a continuous shift between abstract (distant) representations of literary texts and (close) analytical text interpretations.

QuaDrama (On Designing Collaboration in a Mixed-Methods Scenario. Reflecting Quantitative Drama Analytics) utilized German-language plays to analyse the textual and structural properties of dramas. With a focus on character types, the team developed an interface to jointly define and annotate the corpus with the aim of automatically detecting and quantitatively analysing different dramatic character types.

For many projects, a hitherto unsolved issue, vagueness, ambiguity or fuzziness of knowledge triggered the mixing of methods. *DhiMu* (Dhimmis and Muslims – Analysing Multi-Religious Spaces in the Medieval Muslim World) developed an in-

terface allowing tracing and visualizing religious minorities in the medieval Middle East in such a way as not to eliminate the uncertainties concerning source, time and place but to make them useable for interpretation. As such, it enables the integration of formerly marginalized source texts in this context.

ArchiMediaL (Computer Vision and Architectural History at Eye-level: Mixed Methods for Linking Research in the Humanities and in Information Technology) aimed to overcome the absence of meta-information of digitized architectural depictions by using computer vision to recognize architectural image content. When the intended corpus turned out to be unsuitably heterogeneous to train the algorithms, the team came up with a tool to crowdsource serviceable image sets instead.

BachBeatles (Musical Schemata: Modelling Challenges and Pattern Finding) set out to find and model the characteristics of voice-leading schemata present in western music. For automated pattern recognition, two key challenges had to be overcome: the polyphonic structure of music as opposed to the sequential structure of text, and the highly flexible nature of these patterns, as the structural notes in individual voices can be elaborated in very different ways.

The interplay from text and rhythm is at the core of project *Rhythmicalizer* (Free Verse Prosodies: Identifying and Classifying Spoken Poetry Using Literary and Computational Perspectives). It synthesizes the rhythmical features of modern poetry by aligning the written text with the speech of the poet. The visualizations found a means to classify many poems along a fluency/disfluency continuum and provided insight that was applied to question established categories and to develop teaching units at the high school level.

The communication about climate change inspired the project *ANCI* (Interpreting Climate Images on the Internet: Mixing Algorithmic and Interpretive Views to Enable an Intercultural Comparison). It analysed the means to visually communicate climate change depending on communication and cultural context. Its team traced its learning curve as a shift from an interpretative towards a structural view of the images in which pictoriality came to be understood as the interplay of images at various levels.

The interpretation of historic texts was a challenge for *Handwriting* (Detecting Authorship, Hands, and Corrections in Historical Manuscripts. A Mixedmethods Approach Towards the Unpublished Writings of an 18th Century Czech Emigré Community in Berlin), in which the detection of author and scribe, cultural background and language sought for the cross-breeding of text- and image-recognition. The resulting open-source software tool not only helps to unveil the interconnected history of an eighteenth-century Czech émigré community in Berlin but also promises applicability beyond a single language and script.

The project *HerCoRe* (Encoding, Processing and Interpreting Vagueness and Uncertainty in Historical Texts – A Pilot Study Based on Multilingual 18th Century

Texts of Dimitrie Cantemir) addressed the vagueness of language in processes of translation between Latin, Romanian and German in eighteenth century. Aiming to develop a tool for digital text analysis beyond keyword search and statistical information, the work done on the required ontology proved crucial and it currently outshines the ongoing research on annotation. It mirrors back onto the underlying hermeneutic research and inspires new approaches to historical studies.

As the results of the projects so far show, the text-related digital humanities seem to be the most advanced at the moment for various reasons and are developing a variety of approaches and questions. As already thematized above, however, the mixing of methods tapped into further systems of representation of culture such as images, architecture, spoken presentation and music. It touched on relevant questions: How to define the epistemological challenges and opportunities when combining established hermeneutic and sometimes experimental digital approaches in research on literary studies, art history, musicology, and related fields? What are the theoretical and methodological principles across all disciplinary digital approaches? How is it possible to work around conceptual roadblocks towards not only a joint result but also an actually joint understanding of issues at hand in digital humanities?

This volume focuses on driving innovation and conceptualizing the Humanities in the 21st century. It serves as a useful tool for designing cutting-edge research that goes beyond conventional strategies. Its aim is to move beyond the simplifying concepts of 'real science', hard and soft data or qualitative and quantitative analysis. Digital humanities challenge traditional conceptions of research objects and approaches. At some points knowledge was generated to which one of the involved disciplines could not "formulate a question to which this knowledge would be an answer,"[13] because the knowledge simply could not be processed through the respective discourse and the result might not be what the researchers expected. This discomfiture as well as destabilization of habitual grounds is peculiar to digital humanities projects and is precisely what is to be highlighted here as a starting point for creativity and insight.

13 This idea is taken from Claus Pias who wrote: "[…] plötzlich [entsteht] durch Geräte ein Wissen von Bildern, zu dem die Kunstgeschichte keine Frage formulieren kann, auf die dieses Wissen eine Antwort wäre, ein Wissen, das einfach vom kunsthistorischen Diskurs nicht verarbeitbar ist." Claus Pias, "Maschinen/lesbar. Darstellung und Deutung mit Computern", in: Matthias Bruhn (ed.): Darstellung und Deutung in der Kunstgeschichte (visual intelligence, Bd.1) (Weimar: Verlag und Datenbank für Geisteswissenschaften, 2000) 125–144.

Bibliography

Berry, David M. and Anders Fagerjord. *Digital Humanities: Knowledge and Critique in a Digital Age*. Cambridge: Polity, 2017.

Burdick, Anne, Johanna Drucker, Peter Lunenfeld, Todd Presner and Jeffrey Schnapp. *Digital_Humanities.* Cambridge: MIT Press, 2016.

Ciula, Arianna, Øyvind Eide, Cristina Marras, and Patrick Sahle. "Models and Modelling between Digital and Humanities. Remarks from a Multidisciplinary Perspective". In *Historical Social Research / Historische Sozialforschung* 43, no. 4, 2018, 343–61.

Creswell, John W. *Qualitative, quantitative and mixed methods approaches*. Thousand Oaks, Calif.: SAGE, 2003.

Drucker, Johanna. "Humanities Approaches to Graphical Display". In *DHQ*, 2011, 5, 1.

Fam, Dena and Michael O'Rourke (ed.). *Interdisciplinary and Transdisciplinary Failures. Lessons Learned from Cautionary Tales*. Abingdon, Oxon; New York, NY: Routledge, 2020.

Feyerabend, Paul. *Against Method. Outline of an anarchistic theory of knowledge*, [1970]. London et al.: Verso, 2010.

Fleck, Ludwik. *Genesis and Development of a Scientific Fact*. Translated by Frederick Bradley and Thaddeus J. Trenn. Chicago, University of Chicago Press, 1981.

Foucault, Michel. "The Discourse of Language". Translated by Rupert Swyer in *Social Science Information* 10/2, 1971, 7–30.

Heppler, Jason. *What Is Digital Humanities*, 2015, https://whatisdigitalhumanities.com/ [accessed: 17.07.2021].

Kleymann, Rabea: „Datendiffraktion: Von Mixed zu Entangled Methods in den Digital Humanities", In *Fabrikation von Erkenntnis – Experimente in den Digital Humanities* (= Zeitschrift für digitale Geisteswissenschaften / Sonderbände, 5), edited by von Manuel Burghardt, Lisa Dieckmann, Timo Steyer, Peer Trilcke, Niels Walkowski, Joëlle Weis and Ulrike Wuttke, text/html Format, DOI: 10.17175/sb005_008. Wolfenbüttel, 2022.

Krämer, Sybille. „Der 'Stachel des Digitalen' – ein Anreiz zur Selbstreflexion in den Geisteswissenschaften? Ein philosophischer Kommentar zu den Digital Humanities in neun Thesen", position paper of the keynote speech at the annual conference DHD 2018 at Cologne University.

Kuckartz, Udo. *Mixed Methods. Methodologie, Forschungsdesigns und Analyseverfahren* (Mixed Methods. Methodology, research designs and analytical procedures). Springer, Wiesbaden 2014.

Mager, Tino and Carola Hein. "Mathematics and/as Humanities: Linking Humanistic Historical to Quantitative Approaches", In *The Mathematics of Urban Morphol-*

ogy, edited by Luca D'Acci, text/html Format, DOI https://doi.org/10.1007/978-3-030-12381-9_27. Berlin: Springer, 2019.

Mitchell, William J. *The Reconfigured Eye. Visual Truth in the Post-Photograph Era*. Cambridge, Mass. et al.: MIT Press, 1992.

Pias, Claus. „Maschinen/lesbar. Darstellung und Deutung mit Computern", In *Darstellung und Deutung in der Kunstgeschichte* (visual intelligence, Bd.1), edited by Matthias Bruhn, 125–144. Weimar, Verlag und Datenbank für Geisteswissenschaften 2000.

Piotrowski, Michael. "Digital Humanities: An Explication", In *Informatik und die Digital Humanities – Im Spannungsfeld zwischen Tool-Building und Forschung auf Augenhöhe*, edited by Manuel Burghardt and Claudia Müller-Birn, C., DOI: https://doi.org/10.18420/infdh2018-07. Bonn: Gesellschaft für Informatik e.V, 2018.

Schwandt, Silke (ed.). *Digital Methods in the Humanities. Challenges, Ideas, Perspectives*. Bielefeld: Bielefeld University Press 2020.

Terras, Melissa. *Quantifying Digital Humanities, UCL Centre for Digital Humanities*, 2011. https://www.ucl.ac.uk/infostudies/melissa-terras/DigitalHumanitiesInfographic.pdf [accessed: 7.12.2023].

Vanhoutte, Edward. "The Gates of Hell: History and Definition of Digital | Humanities | Computing", In *Defining Digital Humanities: A Reader* edited by Melissa M. Terras, Julianne Nyhan and Edward Vanhoutte, 119–156. London, Routledge, 2016.

Mixed Methods and the Digital Humanities

Andrew Prescott

The Digital Humanities was announced as a field in 2004 with the publication by Blackwell of *A Companion to the Digital Humanities* edited by three distinguished pioneers of digital methods in humanities scholarship, Susan Schreibman, Ray Siemens and John Unsworth.[1] Although the term digital humanities was introduced for the first time in the *Companion*, the work surveyed was the product of scholarly projects and experimentation stretching back more than 50 years. With the advent of the World Wide Web in the 1990s, the range of this work considerably expanded, justifying the introduction of the new term 'Digital Humanities' but, the editors of the *Companion* insisted, the aim remained the same: "using information technology to illuminate the human record, and bringing an understanding of the human record to bear on the development and use of information technology."[2]

The editors of the Companion declared that its publication marked a turning point not only because of its use of the neologism Digital Humanities but also because for the first time a wide range of practitioners and specialists "have been brought together to consider digital humanities as a discipline in its own right, as well as to reflect on how it relates to areas of traditional humanities scholarship."[3] The Companion illustrated the wide range of disciplines exploring the possibilities offered by new digital tools and methods, including the study of literature, linguistics, history, archaeology, classics, lexicography, music, the performing arts and even philosophy and religion. Many of the methodological issues raised by the adoption of digital methods cut across disciplinary boundaries, and the Companion revealed many areas of shared concern across disciplines. In contrast to much conventional humanities work, methodological concerns loom large in the digital humanities, sometimes at the expense of the wider critical and interpretative issues. This common methodological core is one component uniting the digital humanities as a field.

---

1 Susan Schreibman, Ray Siemens and John Unsworth, *A Companion to Digital Humanities* (Oxford: Blackwell, 2004).

2 Schreibman, *Companion to Digital Humanities*, xxiii.

3 Schreibman, *Companion to Digital Humanities*, xxiii.

The present volume documents an impressive range of projects undertaken under the programme 'Mixed Methods in the Humanities' funded by the Volkswagen Foundation. The aim of the Volkswagen Foundation programme harks back to themes evident in the 2004 Companion to the Digital Humanities. The projects supported by the Volkswagen Foundation programme addressed, on the one hand, major interpretative and critical themes in the humanities while, on the other hand, they sought to generate innovative computing methods and solutions. The Volkswagen Foundation programme confronted a tantalizing research question: how can the attempt to address qualitative-hermeneutical issues in the humanities drive forward innovative research in computer science? Achieving a balance between these two imperatives is a dilemma which frequently occurs in the digital humanities. The aims of the Volkswagen Foundation programme were explicitly mixed—the intention was as much to generate technological innovation as to address major issues in the humanities. Arguably a project might have been considered a success if it stimulated new discoveries in computer science but failed completely in its humanities research.

This 'mixed methods' vision, drawn from the ways in which Social Science research sometimes mixes quantitative and qualitative methods, is in many ways central to the digital humanities enterprise. It will be argued here that the idea of 'mixed methods' is a useful template in developing the digital humanities. This chapter will explore ways in which mixed methods fit into the wider concept of digital humanities and consider the extent to which the development of digital humanities since the middle of the 20th century has been an exploration of mixed methods.

Traditionally, the humanities is seen as comprising such well-established disciplines as literature, history, classics and philology with more recent additions such as media and cultural studies. However, in the past 50 years a number of thematic approaches to the humanities have developed which use interdisciplinary and transdisciplinary perspectives to build new links across subject areas, often reaching beyond the humanities. They include medical humanities, health humanities, environmental humanities, public humanities, spatial humanities and bio-humanities. Digital humanities can be seen as one of these thematic flavours of the humanities. Like digital humanities, many of these thematic humanities only crystallized quite recently although work had been going on in the area for some decades. For example, the roots of the environmental humanities can be traced back more than a century, but as an increasing quantity of research at the intersection of history, indigenous studies, anthropology, philosophy, political theory and nonfiction writing emerged, a group of Australian researchers gave it the name 'ecological humanities'. By 2010, the wider term 'environmental humanities' had been widely adopted. It is perhaps natural to stress the humanities component in these designations but it is not necessarily so. In the case of medical humanities, for example, the emphasis is on the training of medical practitioners and enhancement of medical practice. It was

for this reason that health humanities, which stressed the contribution humanities could make to well-being, emerged.

An interdisciplinary approach and enthusiasm for the possibilities of mixing methodologies are characteristic of all these thematic approaches to the humanities. These thematic humanities often describe themselves as working 'at the intersection' of a variety of disciplines. They frequently advocate a methodological bricolage. In a keynote address to a recent symposium on mixed methods in the environmental humanities, Sharlene Hesse-Biber has described the contribution that she feels mixed models can make in this field:

> "I like to think of us taking a kind of mixed model approach where we have a set of questions that are talking to one another. Really, we're weaving a tale of how to figure out what's happening out in the environment, how to get all those stakeholders together. It's not kind of melding things together, but weaving and talking. This idea of weaving voices together [is] not so much in competition with one another but with the goal of complex understanding of meaning. I also like the idea of crossing boundaries; weaving and crossing borders is a way to get at the range of issues out there and to try to understand complex ways of dealing wpolitiqueith these issues."[4]

This emphasis on crossing borders and weaving voices together describes very well some of the approaches taken by projects described in this volume, and illustrates how there is a great deal in common between these new thematic humanities such as digital humanities, environmental humanities, and health humanities.

Following the publication of the *Companion* in 2004, the term digital humanities quickly displaced its precursors such as 'Humanities Computing'. Many labs and professional organizations recast themselves as digital humanities organizations. An Alliance of Digital Humanities Organizations was established.[5] The major international conference in the field was renamed Digital Humanities, and now it attracts hundreds of delegates. As digital humanities rapidly became the accepted designation of the field, however, there was an explosion of literature debating its scope, range and nature. Indeed, in the 10 years after the publication of the *Companion* it seemed as if the field would be overwhelmed by discussion about the nature and scope of Digital Humanities. The Digital Humanities Manifesto 2.0, emulating such models as the Vorticist *Blast!* manifestos, saw digital humanities as a disruptive force that would transform the whole concept of the humanities:

---

4 Sharlene Hesse-Biber, "Transcript of Key Note Address on Mixed Methods and Transformative Approaches to Social and Environmental Justice" in: Janet McIntyre-Mills and Norma R. A. Romm (eds.): *Mixed Methods and Cross Disciplinary Research: Towards Cultivating Eco-Systemic Living* (Cham: Springer, 2019), ix.

5 https://adho.org/about [accessed: 08.08.2021].

> "Digital Humanities is not a unified field but an array of convergent practices that explore a universe in which: a) print is no longer the exclusive or the normative medium in which knowledge is produced and/or disseminated; instead, print finds itself absorbed into new, multimedia configurations; and b) digital tools, techniques, and media have altered the production and dissemination of knowledge in the arts, human and social sciences."[6]

Such claims prompted further debates as to the level of formal technical knowledge required to practice digital humanities, whether the emphasis should be on building digital tools and resources or on critical frameworks, and how valid criticism driven by quantification or algorithms might be.[7] For some commentators, the adoption of digital methods was a youthful insurgency against a fuddy-duddy and hidebound humanities. Mark Sample notoriously commented that "The Digital Humanities should not be about the digital at all. It's all about innovation and disruption. The digital humanities is really an insurgent humanities."[8] The attacks on digital humanities by such eminent critics as Stanley Fish in 2012 were as much as anything a reaction to the exaggerated claims for the field made by Sample and others.[9] As Ted Underwood has remarked, these criticisms seem now to have subsided.[10] One of the concerns of Fish was that more traditional qualitative forms of humanities method might be sidelined by the rise of quantitative methods. In fact, to some degree Fish's criticisms can be seen as a plea for more mixed methods.

It is surprising that these debates about nature and scope of digital humanities should have become so heated, since earlier work on Humanities Computing had already established clear theoretical frameworks which with hindsight appeared surprisingly far-sighted and robust. One area in which it was clear from an early stage that humanities computing would be different to traditional work in the humanities was the need for the research to be undertaken by teams blending computing, humanities and other professional skills. In this respect, humanities computing was similar to other areas adopting new technologies. The importance of assembling teams with a wide range of specialist and professional skills has for example long

---

6 https://www.humanitiesblast.com/manifesto/Manifesto_V2.pdf [accessed: 04.08.2021].

7 Melissa Terras, Julianne Nyhan and Edward Vanhoutte (eds.), *Defining Digital Humanities; A Reader* (Abingdon: Routledge, 2013).

8 Cited in Patrik Svensson, "Envisioning the Digital Humanities", in: *Digital Humanities Quarterly* 6:1 (2012).

9 Stanley Fish, *Think Again: Contrarian Reflections on Life, Culture, Politics, Religion, Law, and Education* (Princeton and Oxford: Princeton University Press, 2019), 343–56.

10 Ted Underwood, "Digital Humanities as a Semi-Normal Thing", in: Matthew Gold and Laura Klein (eds.), *Debates in the Digital Humanities 2019* (Minneapolis: University of Minnesota Press, 2019), 96–8.

been recognized in medical imaging.[11] How universities, which privilege specialist academic knowledge, accommodate the different blends of skills and knowledge necessary to use new digital methods, is a difficult issue, and one that has still not been satisfactorily resolved in either the humanities or the sciences.

These issues as to what humanities computing might be and what structures should support it were already the subject of lively debate in the academy in the 1990s. In Britain, a key moment in the development of digital humanities took place on the banks of Loch Lomond in Scotland in September 1996, when a meeting was held at the Buchanan Hotel entitled 'Defining Humanities Computing'. Attending the meeting were representatives of three leading universities which had been involved in the Computers in Teaching Initiative established in Britain in the early 1990s.[12] Many of the names are familiar still: Harold Short, Willard McCarty and Marilyn Deegan from King's College London; Christian Kay, Jean Anderson and Ann Gow from the University of Glasgow; and Stuart Lee, Mike Popham and Mike Fraser from the University of Oxford. The Buchanan Hotel meeting was perhaps the nearest thing to a digital humanities summit meeting that has ever taken place in Britain.

Many of the questions debated on the banks of Loch Lomond in 1996 remain current and relevant. How should we define humanities computing theoretically or pragmatically in terms of current practice? Where does humanities computing fit within institutions of higher education? How will computing integrate into standard humanities courses? What should research in humanities computing be about? For some attendees at the 1996 meeting, computing facilitated and supported academic research and the role of humanities computing specialists was analogous to that of lab technicians. For others, particularly Willard McCarty, who has been the most persistent and forceful advocate of this view in Britain, humanities computing—and thus by extension digital humanities—is a field of intellectual endeavour and investigation on a par with more traditional academic disciplines such as history, classics or media studies.

In the course of the discussions in Scotland, Willard drafted the following definition of the field as it appeared to him in 1996:

> "HUMANITIES COMPUTING is an academic field concerned with the application of computing tools to humanities and arts data or their use in the creation of these data. It is methodological in nature and interdisciplinary in scope. It works at the

11 See for example B. Rodriguez, A. Carusi et al., "Human-based Approaches to Pharmacology and Cardiology: an Interdisciplinary and Intersectoral Workshop", in: *EP Eurospace* 18: 9 (2016), 1287–98.

12 I am grateful to Willard McCarty for lending me his papers relating to the 1996 meeting, from which all the following information is taken.

> intersection of computing with the other disciplines and focuses both on the pragmatic issues of how computing assists scholarship and teaching in these disciplines, and on the theoretical problems of shift in perspective brought about by computing. It seeks to define the common ground of techniques and approaches to data, and how scholarly processes may be understood and mechanized. It studies the sociology of knowledge as this is affected by computing as well as the fundamental cognitive problem of how we know what we know.
> Within the institution, humanities computing is manifested in teaching, research, and service. The subject itself is taught, as well as its particular application to another discipline at the invitation of the home department. Practitioners of humanities computing conduct their own research as well as participate by invitation in the projects of others. They take as a basic responsibility collegial service, assisting colleagues in their work and collaborating with them in the training of students."

This is a beautifully crafted working definition which applies as much to the digital humanities today as to the humanities computing of 1996. The emphasis on the sociology of knowledge is just as important in the age of social media and Cambridge Analytica as it was in the early days of the Web. But there is an air of passivity and reticence about the definition. There is a focus on reproducing existing scholarly technique rather than on investigating how digital methods can transform scholarship. It is assumed that computers assist scholarship. The way in which digital humanities has moved on towards more experimental projects that might seek to transcend the limitations of existing scholarly method is amply illustrated by the portfolio of projects financed by the Volkswagen Foundation discussed in this volume. But nevertheless these projects all fall within Willard McCarty's broad 1996 definition of humanities computing as a field concerned with the application of computing tools to humanities and arts data. Like the projects discussed in 1996, they are methodological in nature and interdisciplinary in scope.

Another issue that McCarty's definition skirts around is the methodological emphasis of the digital humanities. While emphasising the importance of methodological discussion in the digital humanities, the Loch Lomond formulation stops short of defining how these methodological concerns shape and characterize the digital humanities. One common feature of digital humanities methodologies is the importance of mixed methods, the connecting thread of the Volkswagen Foundation projects discussed in this volume. The heterogeneous nature of humanities data, with their frequent silences, disjunctures and deceptions, means that they cannot always be convincingly and credibly used in conjunction with quantitative techniques. Exploring the potential and pitfalls of such mixed methods is at the heart of the digital humanities.

Indeed, much of the history of the digital humanities can be seen as a dialectic around such mixed methods—a debate as to the extent to which quantitative meth-

ods can be used in conjunction with more hermeneutic forms of analysis, a wariness as to what are the appropriate boundaries of automated analysis and an anxiety to ensure that the critic's voice ultimately remains a human one. In many ways, viewing the history of humanities computing and then digital humanities as a series of debates and exercises exploring these boundaries between the computational and the manual is a useful perspective from which to view the development of digital humanities.

These anxieties about the most appropriate balance between an automated and manual method were expressed from the time that computers began to appear in the 1960s, as is apparent from three articles in a 1965 issue of the *Journal of Higher Education*. Ephim Fogel, while insisting that he saw computers as potentially valuable to humanities scholars, expressed concern as to how the lone humanities scholar would have the time and resources to deploy computing in research or be able to amass sufficient data for quantifiable analysis.[13] Allan Markman, while envisaging that one day all the literature of the world would be available for searching on magnetic tape, nevertheless emphasized that the machine was only a tool: "Man made the machine, and men will use it as a tool. By itself it is nothing."[14] All these were concerns about how far mixed methods could be implemented in literary studies. The historian Franklin Pegues was cautiously optimistic about the assistance computers could provide but insisted that machines must be kept in their place:

> "The machine must serve to free the humanist from time-consuming labor and enlarge his horizons for greater and more important accomplishments. In short, the machine can help the scholar be a better humanist. It will be a sad day for the humanities if scholars seek to undertake only that work which can be computer-oriented."[15]

This assumption that the machine must only ever be a tool and subservient to the higher aims of humanities scholarship remains common. Again, it can be interpreted as nervousness about the nature and extent to which the use of digital technology in humanities scholarship promotes and requires the use of mixed methods.

The same tension between quantification and hermeneutics as humanities scholars began to explore the potential of computers can be seen in a landmark article by the archaeologist David Clarke published in 1973, 'Archaeology: the Loss of

13 Ephim G. Fogel, "The Humanist and the Computer: Vision and Actuality", in: *Journal of Higher Education* 36: 2 (1965), 61–8.

14 Allan Markman, "Litterae ex Machina: Man and Machine in Literary Criticism", in: *Journal of Higher Education* 36:2 (1965), 79.

15 Franklin J. Pegues, "Computer Research in the Humanities", in: *Journal of Higher Education* 36:2 (1965), 107.

Innocence'.[16] Clarke described how changes in the technological environment as a result of the Second World War had fostered the professionalization of archaeology. Among the new possibilities which had precipitated that process was the computer which, in Clarke's words, "provides an expanding armoury of analogue and digital techniques for computation, experimentation, simulation, visual display and graphic analysis … They also provide powerful hammer-and-anvil procedures to beat out archaeological theory from intransigent data".[17] Clarke went on to observe that "A major embarrassment of the computer has been that it enabled us to do explicitly things which we had always claimed to do intuitively. In many cases it is now painfully clear that we were not only deceived by the intuitions of innocence but that many of the things we wished to do could not be done or were not worth doing and many unimagined things can be done".[18] The Processual Archaeology or New Archaeology advocated by Clarke placed computer analysis at the heart of a rigorous scientific approach to archaeological data. But the overemphasis of processual archaeology on quantification brought charges of ignoring human agency. Computers became associated in Archaeology with number-crunching, and it is only more recently with the rise of new non-quantitative techniques such as 3D imaging and GIS that computers have again emerged as key archaeological tools.[19] Similar anxieties as a result of the false assumption that computers were chiefly about counting were also evident at the same period in discussions in the historical discipline, particularly in the debates around cliometrics.[20]

These examples illustrate the varied roots of the digital humanities. It is unfortunate that the growth of the digital humanities is often presented in an over-simplified form. The origins of the digital humanities are usually linked to the pioneering work of Father Roberto Busa who in the 1940s collaborated with IBM to create a massive computerized index to the works of St Thomas Aquinas. There is no doubting the scale and influence of Busa's achievement, using first punch card and then magnetic tape to process 22 million words, finally producing the 20 million lines and 65,000 pages of the Index Thomisticus in 1980. Busa's claim to be the founding father of the digital humanities was reinforced by the fact that he provided a foreword to the 2004 Companion. Question marks have been raised against Busa's legacy recently—it has been pointed out that many of those who entered Busa's data were

16 David Clarke, "Archaeology: the Loss of Innocence", in: *Antiquity* 47: 185 (1973), 6–18.

17 Clarke, "Loss of Innocence", 9.

18 Clarke, "Loss of Innocence", 10.

19 Bruce Trigger, '"The Loss of Innocence" in Historical Perspective', n: *Antiquity* 72:277 (1998), 694–8; Jeremy Huggett, "The Past in Bits: Towards an Archaeology of Information Technology?", in: *Internet Archaeology* 15 (2004): https://doi.org/10.11141/ia.15.4.

20 M. Haupert, "The Impact of Cliometrics on Economics and History", in: *Revue d'économie politique* 127: 6 (2017), 1059–1082.

anonymous forgotten women, and the fact that IBM had in the 1940s collaborated with fascist regimes in Europe has recently been pointed out.[21]

The main problem with the Busa digital humanities foundation myth, however, is the way in which it restricts perspectives on the development of the digital humanities to a single line of development which privileges the encoding and manipulation of text. There are many other routes of development of digital humanities which have been given less prominence. If more attention were paid to these other paths of development of digital humanities, a richer view of the range and potential of digital humanities would emerge, and some of the circular debates about the nature of the digital humanities and its place in scholarship might be avoided.

For example, the roots of many of the large digital packages most widely used by humanities researchers in the English-speaking world lay in the library and commercial world and followed a different path of development. Libraries quickly realized the potential of computing to streamline cataloguing work, particularly by sharing records compiled to common bibliographic standards. A pioneer of this work was Henrietta Avram (1919–2006), a computer specialist at the Library of Congress, who in the 1960s developed the standard for automated library catalogue records, the Machine Readable Catalogue (MARC).[22] MARC was adopted in the 1970s for large-scale bibliographical enterprises such as the Eighteenth Century Short Title Catalogue which afterwards became the English Short Title Catalogue (ESTC), listing publications in Britain and North America from 1473 to 1800. The availability of the ESTC MARC records both online and in CD ROM from the 1980s enabled many new lines of research into the printed record of the early modern period to be more readily pursued, such as commercial partnerships in printing or the geographical distribution of printers.

While libraries were increasingly able to share and distribute catalogue information in printed form, great strides were also being made in distributing images of

21 M. Terras and J. Nyhan, "Father Busa's Female Punch Card Operatives", in: M. Gold (ed.): *Debates in the Digital Humanities*, 60–65 (Minneapolis, London: University of Minnesota Press, 2016). Steven E. Jones, *Roberto Busa S.J. and the Emergence of Humanities Computing: The Priest and the Punched* Cards (New York and London: Routledge, 2016).; A. Jacob, "Punching Holes in the International Busa Machine Narrative", Interdisciplinary Digital Engagement in the Arts and Humanities (IDEAH), 1:1 (2020): https://ideah.pubpub.org/pub/7yvuobes [accessed: 05.08.2021].

22 For the following, see Stephen Gregg, *Old Books and Digital Publishing: Eighteenth-Century Collections Online* (Cambridge: Cambridge University Press, 2020).; A. Prescott, "Searching for Dr Johnson: The Digitization of the Burney Newspaper Collection", in: S. G. Brandtzæg, P. Goring, and C. Watson (eds.), *Travelling Chronicles: News and Newspapers from the Early Modern Period to the Eighteenth Century* (Leiden: Brill, 2018), 49–71; and A. Prescott and L. Hughes, "Why Do We Digitise: the Case for Slow Digitisation", in: *Archive Journal*, September 2018, available at: https://www.archivejournal.net/essays/why-do-we-digitize-the-case-for-slow-digitization/.

the books and manuscripts held by libraries and archives. As early as 1938, H.G. Wells envisaged a world brain consisting of a universal library on microfilm. The British Museum became interested in the extent to which microfilming could reduce wear and tear on its collections and in 1935 Eugene Power, who afterwards established University Microfilms International which pioneered the large-scale microfilm publication of primary materials, helped set up a programme at the British Museum for the microfilming of rare books. By the 1970s, microform publication was a major commercial activity and firms such as Universal Microfilms, Primary Sources Microfilm (an imprint of Gale) and Chadwyck-Healey vied to maximize their coverage of such major bibliographic resources as the ESTC.

When digital imaging became more widespread in the 1980s, it was an obvious step to amalgamate scans of microfilm with bibliographic records. Thus, Early English Books Online (EEBO) combined ESTC records for 1473–1700 with digitized images from microfilm to provide users with access to the corpus of English language printing for this period. Subsequent projects went further, with Eighteenth Century Collections Online (ECCO) and the Burney Newspapers applying optical character recognition to the microfilm scans to try and produce fully searchable packages. The poor quality of the microfilm scans and the difficulties of automatically converting archaic letter forms mean that the accuracy of the searches performed in these packages is very low, but nevertheless these large-scale commercial packages, covering great swathes of the primary sources used by many humanities disciplines, have established themselves as indispensable tools.

Packages such as EEBO and ECCO represent a separate path of development in the digital humanities to the Busa foundation myth, one rooted in the adoption of computing in libraries and the introduction of imaging technologies such as microform. There is also a much stronger commercial element in this line of development, notwithstanding Busa's collaboration with IBM. The key players in bringing EEBO and ECCO to fruition were the microform publishers ProQuest (the successor to University Microfilms which also bought Chadwyck-Healey) and Gale. Moreover, the influence of these pioneering projects can also be seen in subsequent developments such as Google Books and the tools derived from it. Although digital imaging is a vast improvement on microfilm, not least because of its ability to readily provide colour images, much of our current use of digital imaging is dependent on methods and procedures established in large-scale microfilming projects in the 1930s.

In a similar fashion, the use of specialist imaging techniques such as multi-spectral imaging to examine historic records and recover faded text owes much to the pioneering work undertaken using ultraviolet and infrared photography in the first half of the twentieth century. The Benedictine monk Rudolph Kögel demonstrated

in 1914 how ultraviolet fluorescence could be used to recover damaged text.[23] The Chaucer scholar John Manly presented an ultraviolet lamp to the British Museum in 1929,[24] and the roots of the forensic imaging, which has characterized many digital humanities projects ranging from the Electronic Beowulf to the recovery of the field diaries of the Victorian missionary David Livingstone, can be found in these early experiments.[25]

Another major source which has appeared in the digital humanities but has also been overshadowed by the Busa legend is the use of databases in historical computing. Manfred Thaller, himself one of the great pioneers of the use of computing in history, has lamented how the way in which the early use of computing by historians relied on relational databases has been forgotten.[26] The punched cards which Busa used had been developed to enable automated processing of one of the great historical information sources, the census. Herman Hollerith pioneered the use of automated sorting of punch cards to expedite data handling in the 1890 US Census.[27] Historians quickly recognized the potential of methods used to deal with large amounts of data such as that generated by the census. From the mid-1950s, economic historians in the United States such as Robert Fogel, Douglass North and Lance Davies began to use computers to analyse more precisely topics such as the economics of slavery.[28] This encouraged British scholars such as Roderick Floud to use computing in investigating issues such as historic standards of living.[29] In France, Lucien Febvre, one of the founders of the Annales school of history, dreamt of historical laboratories similar to scientific ones, and this vision came closer to reality when historians such as Michel Vovelle used computers in the 1960s and 1970s to analyse thousands of wills.[30]

The German quantitative historian and computer scientist, Manfred Thaller, in an interview with Julianne Nyhan gives a vivid account of the difficulties of

23 Barry Knight, "Father Kögel and the Ultra-violet Examination of Manuscripts", 2014. https://blogs.bl.uk/collectioncare/2014/03/father-k%C3%B6gel-and-the-ultra-violet-examination-of-manuscripts.html [accessed: 08.08.2020].

24 Prescott and Hughes, "Slow Digitisation".

25 ebeowulf.uky.edu [accessed 08.08.2021]; livengstone.online.org [accessed 08.08.2021].

26 M. Thaller, "On Vagueness and Uncertainty in Historical Data", 2020. https://ivorytower.hypotheses.org/88#more-88 [accessed 08.08.2021].

27 K. Driscoll, "From punched cards to "big data": a social history of data-base populism", in: *Communication +1*, 1 (2012), article 4.

28 https://www.nobelprize.org/prizes/economic-sciences/1993/fogel/biographical/ [accessed 08.08.2021].

29 https://archives.history.ac.uk/makinghistory/resources/interviews/Floud_Roderick.html [accessed 08.08.2021].

30 Michelle Volvelle, *Piété baroque et déchristianisation en Provence au XVIIIe siècle : Les attitudes devant la mort d'après les clauses des testaments* (Paris : Plon, 1973).

developing databases for historical research in these early days.[31] There was no consensus about what tools and languages were most appropriate and generally the only way forward was to learn the programming language for oneself. Thaller began to imagine a package comparable to those used in Social Sciences which would be specifically geared to the difficulties of processing historical sources. It would enable the user to capture all the different features of a particular historical source and would also cope with the inconsistencies, gaps and uncertainties of historical sources. This historical software system, CLIO, was developed with support from the Volkswagen Foundation and became enormously influential in the 1980s and early 1990s. Although it has fallen from favour with the growth of the PC and generalized packages such as Microsoft Office, nevertheless there is a great deal of value in the way in which CLIO approaches historical sources. Thaller has continued to produce many other pioneering projects such as the digitization of the Duderstadt Municipal Archive, also funded by the Volkswagen Foundation, and in his retirement Professor Thaller is exploring how computers can be redesigned so as to better represent historical sources—a truly visionary project.[32]

The roots of the Digital Humanities thus spread far and wide, considerably beyond the narrow scope implied by the suggestion that the field emerged from beneath the cloak of Father Busa. The work of film historians and pioneers of oral history also fed into the development of digital techniques. Research into the digital processing of sound and music has been enormously influential in the development of the digital humanities. Creative practice has been another major source of inspiration. Already by the 1950s, artists were experimenting with the creative possibilities of devices such as oscilloscope, plotter and early animation. In 1966, a group of contemporary artists joined forces with computer scientists and engineers from Bell Labs in the United States to host a series of performances using new technologies. A pioneering exhibition of algorithmic art called 'Cybernetic Serendipity' was held at the Institute of Contemporary Arts in 1968.[33] This included early electronic musical instruments by pioneers such as Peter Zinoviev. The origins of the digital humanities lie as much in the work of electronic composers such as Daphne Oram,[34] who helped

31 Manfred Thaller and Julianne Nyham, "It's Probably the Only Modestly Widely Used System with a Command Language in Latin: Manfred Thaller and Julianne Nyhan", 2014. https://hiddenhistories.github.io/manfred-thaller [accessed 08.08.2021]. See also M. Thaller, "Between the Chairs: An Interdisciplinary Career", in: *Historical Social Research*, Supplement, 29 (2017), 7–109, https://doi.org/10.12759/hsr.suppl.29.2017.7-109.

32 https://ivorytower.hypotheses.org/author/mata [accessed 09.08.2021].

33 https://archive.ica.art/whats-on/cybernetic-serendipity-documentation [accessed 09.08.2021].

34 https://www.daphneoram.org/ It is worth also looking at the 2021 film by Lisa Rovner, *Sisters with Transistors:* https://sisterswithtransistors.com/ [accessed 09.08.2021].

create the BBC's Radiophonics Workshop in the 1950s, as they do in the heritage of Roberto Busa.

Once we understand the diverse roots of the digital arts and humanities, their multi-faceted and varied character makes more sense. There is no single discipline or clear-cut set of methods in the digital humanities; it is a creative and intellectual ecosystem. It is not surprising that this is a world of mixed methods.

There is no space here nor would it be practical to discuss how mixed methods might be deployed in future in the digital humanities and what this indicates for the development of the field. The various project reports assembled here illustrate some of the possibilities. Given the scope of the Digital Humanities, it is not surprising that these are wide-ranging. The ways in which digitization gives us new ways of viewing text inevitably loom large. The ways in which distant readings of text might combine both quantitative and qualitative insights is a rich area of investigation. It is also intriguing to consider how theoretical concepts from the humanities such as the idea of intertextuality can be measured and documented. Some of the most exciting challenges in using mixed methods in the digital humanities lie in multimedia—in developing ways of quantifying features of the vast image banks that have now been created or in exploring digital music. Reading over methodological insights from one area to the other remains one of the most exciting areas of development for Digital Humanities.

But perhaps the most tantalizing area for further research is that summarized by Manfred Thaller as follows:

> "Historians face a specific challenge: they need to derive conclusions from evidence which is always incomplete, contradictory and anything but precise."[35]

While Thaller insists on the importance of disciplinary focus, and is uneasy about the way in which the idea of digital humanities weakens these digital boundaries, it can be argued that this problem is one that occurs in many humanities disciplines—the primary source materials on which we rely are often vague, contradictory and full of gaps. It challenges many of the ideas of information which have underpinned ideas of computing since the time of Claude Shannon. In approaching these inconsistent and frequently misleading historical data, we need mixed methods, and it is for this reason that their potential, so well illustrated in this volume, is an exciting avenue for the digital humanities.

35 Thaller, "Vagueness and Uncertainty".

Bibliography

"Daphne Oram". https://www.daphneoram.org/. Last modified 9 August, 2021.
"Making History". Last modified 8 August, 2021 https://archives.history.ac.uk/makinghistory/resources/interviews/Floud_Roderick.html [accessed 08.08.2021].
Alliance of Digital Humanities Organization. Last modified 8 August, 2021. https://adho.org/about.
Clarke, David. "Archaeology: the Loss of Innocence". In Antiquity 47: 185 (1973), 6–18.
Driscoll, Kevin. "From punched cards to "big data": a social history of data-base populism", in: Communication +1, 1 (2012), article 4.
ebeowulf.uky.edu [accessed 08.08.2021].
Fish, Stanley. Think Again: Contrarian Reflections on Life, Culture, Politics, Religion, Law, and Education. Princeton and Oxford: Princeton University Press, 2019.
Fogel, Ephim G. "The Humanist and the Computer: Vision and Actuality". In Journal of Higher Education 36: 2 (1965), 61–68.
Gregg, Stephen. Old Books and Digital Publishing: Eighteenth-Century Collections Online. Cambridge: Cambridge University Press, 2020.
Haupert, Michael. "The Impact of Cliometrics on Economics and History", in: Revue d'économie politique 127: 6 (2017), 1059–1082.
Hesse-Biber, Sharlene. "Transcript of Key Note Address on Mixed Methods and Transformative Approaches to Social and Environmental Justice". In Mixed Methods and Cross Disciplinary Research: Towards Cultivating Eco-Systemic Living, edited by Janet McIntyre-Mills and Norma R. A. Romm, vii-x. Cham: Springer, 2019.
https://www.humanitiesblast.com/manifesto/Manifesto_V2.pdf [accessed 04.08.2021].
https://www.nobelprize.org/prizes/economic-sciences/1993/fogel/biographical/ [accessed 08.08.2021].
Huggett, Jeremy. "The Past in Bits: Towards an Archaeology of Information Technology?". In Internet Archaeology 15 (2004): https://doi.org/10.11141/ia.15.4.
Institute of Contemporary Art. Last modified 9 August, 2021. https://archive.ica.art/whats-on/cybernetic-serendipity-documentation.
Jacob, Arum. "Punching Holes in the International Busa Machine Narrative", In Interdisciplinary Digital Engagement in the Arts and Humanities (IDEAH), 1:1 (2020): https://ideah.pubpub.org/pub/7yvuobes [accessed 05.08.2021].
Jones, Steven E. Roberto Busa S.J. and the Emergence of Humanities Computing: The Priest and the Punched Cards. New York and London: Routledge, 2016)
Knight, Barry. "Father Kögel and the Ultra-violet Examination of Manuscripts", 2014. https://blogs.bl.uk/collectioncare/2014/03/father-k%C3%B6gel-and-the-ultra-violet-examination-of-manuscripts.html [accessed 08.08.2020].

Lisa Rovner, Sisters with Transistors: https://sisterswithtransistors.com/. Last modified 9 August, 2021.

livengstone.online.org [accessed 08.08.2021].

Markman, Allan. "Litterae ex Machina: Man and Machine in Literary Criticism". In Journal of Higher Education 36:2 (1965), 69–79.

Pegues, Franklin J. "Computer Research in the Humanities". In Journal of Higher Education 36:2 (1965), 105–108.

Prescott, Andrew, and Lorna M. Hughes. "Why Do We Digitise: the Case for Slow Digitisation". In Archive Journal, September 2018, available at: https://www.archivejournal.net/essays/why-do-we-digitize-the-case-for-slow-digitization/.

Prescott, Andrew. "Searching for Dr Johnson: The Digitization of the Burney Newspaper Collection". In Travelling Chronicles: News and Newspapers from the Early Modern Period to the Eighteenth Century, edited by Siv Gøril Brandtzæg, Paul Goring and Christine Watson, 49–71. Leiden: Brill 2018.

Rodriguez, Bianca, Annamaria Carusi et al. "Human-based Approaches to Pharmacology and Cardiology: an Interdisciplinary and Intersectoral Workshop". In EP Eurospace 18: 9 (2016), 1287–1298.

Schreibman, Susan, Ray Siemens, and John Unsworth. A Companion to Digital Humanities. Oxford: Blackwell, 2004.

Svensson, Patrik. "Envisioning the Digital Humanities". In Digital Humanities Quarterly 6:1 (2012).

Terras, Melissa, and Juianne Nyhan. "Father Busa's Female Punch Card Operatives". In Debates in the Digital Humanities, edited by Matthew K. Gold, 60–65. Minneapolis, London: University of Minnesota Press, 2016.

Terras, Melissa, Julianne Nyhan, and Edward Vanhoutte (eds.). Defining Digital Humanities. A Reader. Abingdon: Routledge, 2013.

Thaller, Manfred, and Julianne Nyham. "It's Probably the Only Modestly Widely Used System with a Command Language in Latin: Manfred Thaller and Julianne Nyhan", 2014. https://hiddenhistories.github.io/manfred-thaller [accessed: 08.08.2021].

Thaller, Manfred. "Between the Chairs: An Interdisciplinary Career". In Historical Social Research, Supplement, 29 (2017), 7–109, https://doi.org/10.12759/hsr.suppl.29.2017.7-109.

Thaller, Manfred. "On Vagueness and Uncertainty in Historical Data", 2020. https://ivorytower.hypotheses.org/88#more-88 [accessed: 08.08.2021].

Thaller, Manfred. A Digital Ivory Tower. Last modified 9 August, 2021. https://ivorytower.hypotheses.org/author/mata.

Trigger, Bruce. "'The Loss of Innocence' in Historical Perspective". In Antiquity 72:277 (1998), 694–698.

Underwood, Ted. “Digital Humanities as a Semi-Normal Thing”. In Debates in the Digital Humanities 2019 edited by Matthew Gold and Laura Klein, 96–100. Minneapolis: University of Minnesota Press, 2019.

Volvelle, Michelle. Piété baroque et déchristianisation en Provence au XVIIIe siècle: Les attitudes devant la mort d’après les clauses des testaments. Paris: Plon, 1973.

II TEN CASE STUDIES

#PARAPHRASE

#Paraphrase denotes, in the understanding of the Oxford English dictionary, a "rewording of something written or spoken by someone else, esp. with the aim of making the sense clearer". As such it is a practice of discussing, transforming or even conceptually translating an idea or concept across media such as language and writing, music or, at times, art, with the latter "elaborating on a well-known tune" or visual motive.*

While the term itself remains deeply rooted in the humanities so far, the associated challenge fits well with digital humanities: "The process of paraphrasing can be intended by an author, however the phenomenon is highly dependent on the interpretation of the hearer/reader – if he/she does not detect the similarity and relate the posttext to a pretext, it stays a 'phrase' instead of a 'paraphrase'. The phenomenon is therefore accidental (in the philosophical sense)" (Digital Plato).

In consequence, the detection of paraphrases is part of the complex sphere of intertextuality – the way in which texts within and beyond a specific #Corpus relate to each other – and addresses the sphere between #Similarity and #Uncertainty.

* "paraphrase, n". in: *Oxford English Dictionary (OED)*, Third Edition, June 2005; most recently modified version published online March 2022, https://www.oed.com/ [25.10.2022].

Title: "Digital Plato – Tradition and Reception"
Team: Charlotte Schubert, Roxana Kath, Michaela Rücker (Department of Ancient History, Leipzig), Kurt Sier, Eva Wöckener-Gade (Department of Classical Philology, Leipzig), Paul Molitor, Jörg Ritter, Marcus Pöckelmann (Department of Computer Science, Halle), Joachim Scharloth, Simon Meier-Vieracker, Franz Keilholz, Xiaozhou Yu (Department of Applied Linguistics, Dresden / Waseda University Tokio)
Advisory Board: Jonas Grethlein (University of Heidelberg, Department of Classical Philology), John Nerbonne (University of Groningen, Department of Computational Linguistics), and Manfred Pinkal (University of Saarland, Department of Computational Linguistics)
Corpus: The whole of Ancient Greek literature (TLG-E)
Field of Study: Intertextuality, (Classical) Receptions Studies, Natural Language Processing, Paraphrase Extraction / Paraphrase Search
Institution: Universities of Dresden, Halle, Leipzig, and Waseda University Tokio
Methods: Hermeneutics, word embedding, neural networks, close and distant reading, visualisation, web development
Tools: Paraphrase search via rWMD / via n-grams, Reference-Annotator, URN-generator
Technology: Word Mover's Distance, word2vec, complex n-grams, lemmatization, Ruby, Ruby-on-Rails, Python, JavaScript

Innovation in Loops: Developing Tools and Redefining Theories within the Project 'Digital Plato' (Digital Plato)

Eva Wöckener-Gade, Marcus Pöckelmann

Abstract *One of the great desiderata in the field of Classics unfulfilled for centuries has been to trace the reception of the tremendously influential thinker and writer Plato throughout antiquity. The task is not only challenging because of the vast amount of literature (even if we focus on Greek) but has also been impossible to solve with digital approaches so far, especially because many voices of this reception neither explicitly refer to Plato nor cite his words verbatim. A precondition to further explore these instances was, therefore, to be able to detect them. In an iterative process of discussion, implementation, and evaluation, the project developed two web-based solutions for paraphrase search on the Ancient Greek corpus fit for different research questions (based on close and distant reading) together with different supportive resources and tools. The shaping of the theoretical framework, hand in hand with the work on the algorithms, did not only yield innovative digital methods for searching Ancient Greek texts but also sparked a reconsideration of the literary concepts of 'paraphrase' and 'intertextuality'.*

Project description

The Digital Plato project, led by Charlotte Schubert, Paul Molitor, Jörg Ritter, Joachim Scharloth, and Kurt Sier, aimed to explore the aftermath and reception of Plato in Ancient Greek literature through the paraphrasing tradition of Plato's oeuvre in antiquity. The influence of the outstanding thinker (428/427 – 348/347 BC), who laid the foundation for many scientific disciplines, philosophy being one of them, is not easy to trace or to grasp: his impact on Greek literature is inherent in countless texts and manifests itself in different ways. Based on existing approaches from paraphrase detection and adjacent fields, we have developed an innovative paraphrase search for Ancient Greek with the help of which not only verbatim citations but also non-literal references such as allusions can be detected.

The disciplines involved are Ancient History, Classical Philology, Computer Science, and Corpus Linguistics. The team comprised four working groups, two from

the University of Leipzig, one from the University of Halle, and one from the University of Dresden.

Our research question addressed Greek literature from the whole period of antiquity and into the Byzantine era, and we built our corpus respectively: As our database, we used a cleaned and adapted version of the corpus of Ancient Greek literature based on the Thesaurus Linguae Graecae (TLG, https://stephanus.tlg.uci.edu/, Version E), which includes texts from the period approximately between the 8th c. BC and 16th c. AD. Besides some smaller prerequisites (e.g. a lemmatiser identifying the basic form for any given grammatical form for Ancient Greek and a *CTS-URN*-generator for referencing text passages precisely with unique *Uniform Resource Names*), the tools developed by the project for working with this corpus encompass mainly two types of paraphrase search and the 'Reference-Annotator'. With the paraphrase search via rWMD (relaxed Word Mover's Distance), paraphrase candidates characterised by semantic similarity to a passage defined by the user can be identified. In this paper, we will focus on this method; see below for more information. With the second approach for finding paraphrases, which is based on complex n-grams (i.e. sequences of diverse linguistic components), longer text segments, for example works from different authors, can be searched for paraphrase candidates; here, different paraphrase terms for the respective search queries can be modelled by adjusting various parameters. With the Reference-Annotator, the relations between text passages identified as intertextual references, such as paraphrases, can be annotated, analysed, and categorised according to linguistic and literary criteria; see below for a glimpse of the tool. In our web portal, all text passages are provided with persistent identifiers called CTS-URNs, according to the protocol of Canonical Text Services[1], to ensure a stable citation method on the level of single words (instead of page/chapter/line which would not be precise enough for our means).

By implementing different approaches, we mixed methods taken not only from Computer Science and Computational Linguistics but also from hermeneutics. The paraphrase search via complex n-grams is suitable for distant reading while the rWMD-search and the Reference-Annotator are optimised for close reading. Thus, the user can search for references to a special passage from Plato in which he is interested and analyse the results in detail with the latter two tools; on the other hand, he can compare longer texts, for example, a Platonic dialogue with a later work to detect similarities with the former. Different types of visualisations offer the user an alternative and intuitive approach to the results.

1 Christopher Blackwell and David N. Smith, "The Canonical Text Services protocol", GitHub cite-architecture, last modified Apr 10, 2014. https://github.com/cite-architecture/cts_spec/blob/master/md/specification.md .

Our tools are implemented into a freely accessible web interface specially designed for this purpose (https://paraphrasis.org/). The application is not limited to the domain of Plato's works, but allows searching in an author-independent manner within the corpus. Moreover, not only references but also models can be detected since the similarity of the passages compared is the crucial factor (see below). Since the TLG-E is licensed, we chose to offer a test version with texts from a free corpus (based on the Perseus Digital Library) to all users and a full version to those who hold the licence for the TLG (on request). In addition to this portal, according to the FAIR data principles,[2] we provide further information about our research and access to various project-specific resources on our homepage (https://digital-plato.org/).

The project contributed to a multitude of research questions in different areas. We found and collected previously undetected paraphrases of Plato and other authors—which was obviously one of the main aims of our project. The analysis of these paraphrases and the work with the digital tools have stimulated us to rethink both the concepts of paraphrase and intertextual phenomena in literature. Some representative findings have been published by the project in a monograph with contributions from all disciplines.[3]

Methodical background: how to process the literary phenomena intertextuality and paraphrase

The long-term objective of our project was to disclose the multiple forms of Platonic reception in Ancient Greek literature with the aid of digital tools. While obvious forms such as citations are easily detected both digitally and traditionally, indirect forms like allusions pose greater problems. However, all of them expose a certain degree of similarity with their model; otherwise they could not be discovered without further indication in the context. Where this similarity does not involve exact matching of words (citation), we usually find variations of the original wording (paraphrase). An important part of our methodological reflection was, therefore, concerned with the phenomenon *intertextuality* and with *paraphrase* as a specimen of it, and how (or if) they can be processed, which was crucial for our research question: Developing a substantial understanding of what we were searching for was, of course, a precondition for detecting paraphrases and similar forms of intertextual relations.

2 Mark D. Wilkinson, Michel Dumontier, IJsbrand J Aalbersberg, et al. "The FAIR Guiding Principles for scientific data management and stewardship". *Sci Data* 3, 160018 (2016): n. pag.

3 See Charlotte Schubert, Paul Molitor, Jörg Ritter, Joachim Scharloth, and Kurt Sier (ed.), *Platon Digital: Tradition und Rezeption* (Digital Classics Books, Band 3, Heidelberg: Propylaeum, 2019).

Intertextuality encompasses the various forms of relations between different texts and was first introduced by post-structuralist Julia Kristeva as a concept inherent to any text.[4] It has fuelled numerous innovative works on possible interconnections between texts from a new perspective in the Humanities, while former scholars had focused almost exclusively on the author and his intention. However, the post-structuralist conception of intertextuality was so universal that its application was in danger of becoming arbitrary and soon scientific discussions about a possible definition arose. In their course, scholars specified criteria to enlarge the significance and the applicability of the concept of intertextuality, reintegrating structuralist elements, especially regarding the role of the author, into it. Influential until today is Pfister's methodical approach to measuring the degree of intertextuality by using six criteria.[5] Since many of these criteria depend on the interpretation of the reader, they yield somewhat clearer but still partly biased results and are thus hardly applicable for quantitative analysis, as Pfister himself willingly admitted.[6] Consequently, we could only use them for the interpretation of the results of our paraphrase search in Ancient Greek literature—that is for qualitative insights, but not for their generation.

In contrast to intertextuality, which is rooted in literary studies, *paraphrase* has been investigated mainly as a linguistic phenomenon in the context of Natural Language Processing (NLP). Starting from fairly general definitions as well, for example "approximate conceptual equivalence among outwardly different material",[7] various divergent approaches to systematisations tried to identify th e different possible types of transformations involved in paraphrasing.[8] Contrary to most of the studies on intertextuality, the categorisations and taxonomies resulting from these studies were built with the declared goal of 'operationalisability', that is, to be adapted for quantitative analysis. When we tried to apply these categorisations to the analysis of

4 Julia Kristeva, "Wort, Dialog und Roman bei Bachtin (1967)", in: Jens Ihwe (ed.): *Literaturwissenschaft und Linguistik. Ergebnisse und Perspektiven. Band 3: Zur linguistischen Basis der Literaturwissenschaft II* (Frankfurt am Main: Athenäum, 1972), 345–375.

5 Manfred Pfister, "Konzepte der Intertextualität". In *Intertextualität: Formen, Funktionen, anglistische Fallstudien*, ed. Ulrich Broich and Manfred Pfister (Tübingen: De Gruyter, 1985), 1–30.

6 Pfister 1985, 30.

7 Robert-Alain DeBeaugrande, and Wolfgang U. Dressler, "Einführung in die Textlinguistik", *Konzepte der Sprach- und Literaturwissenschaft 28* (Tübingen: De Gruyter, 1981), 50. For an overview of the different proposals cf. Ho et al 2012, 856 et seqq.: Chuk F. Ho, Masrah A. A. Murad, Shyamala, Doraisamy, and Rabiah A. Kadir, "Extracting lexical and phrasal paraphrases: a review of the literature", *Artificial Intelligence Review*, 42.4 (2012), 851–894.

8 Marta Vila, M. Antonia Martí, and Horacio Rodríguez, "Is This a Paraphrase? What Kind? Paraphrase Boundaries and Typology", *Open Journal of Modern Linguistics* 4 (2014), 205–218, here 208–209, 211–212; Rahul Bhagat, and Eduard Hovy, "What Is a Paraphrase?", *Computational Linguistics* 39 (2013), 463–472.

our gold standard set of paraphrases, which we had built for the statistical evaluation of all our approaches, we found that many of these supposedly 'hard' categories still left much room for interpretation, such as, how exactly a synonym should be defined and by whom. Questions like this one are far from being trivial for scholars working with literature in a foreign language from another time and culture with no native speakers to interview. Also, answering such questions is crucial for the generation of resources needed in NLP.

Moreover, many instances of paraphrases from our (in every imaginable way heterogeneous) corpus of Ancient Greek texts still proved to be far too complex to be described within the limits of these (already quite refined) categorisations. One obstacle is posed by the fact that they were developed on the basis of and for the performance on large, rather homogenous corpora of modern English. Still, the limited applicability would probably hold true for many other corpora encompassing literary texts regardless of the language or period. Regarding both problems, a more profound exchange between the disciplines is called for in which insights from both fields, NLP and the humanities, are taken into account and discrepancies are not marginalised or glossed over to achieve applicability of existing approaches. At least for our team, a fresh look at the problem and an intensified dialogue between the disciplines were fruitful to overcome these obstacles. In the course of our work, we achieved better results by applying 'softer' categories not only to the analysis of paraphrases but also as the basis of the tools developed to detect them.

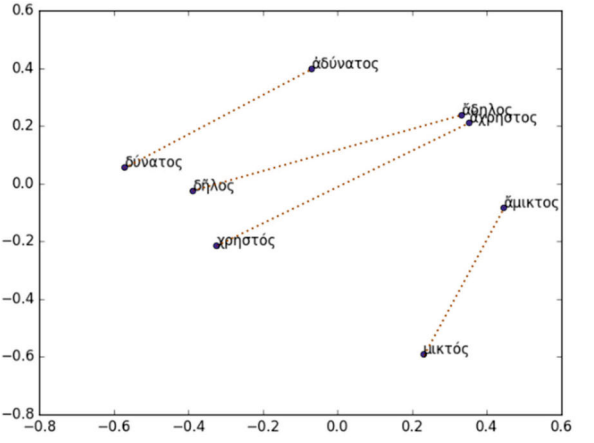
Two loops: old problems lead to new digital methods that, in turn, lead to new concepts

The incompatibility of the existing scientific approaches from both fields—literary studies and linguistics—with each other and with our research question had practical and theoretical consequences. First, we decided not to try to transfer too specific rules from other systems to ours but instead to adapt an approach rooted in very essential assumptions. This is particularly evident in our newly developed rWMD-search which is described in greater detail below. Based on the distributional hypothesis[9] that words occurring in the same contexts tend to have similar meanings, a multidimensional vector space has been built by us representing the text corpus and all its words: a specific vector is assigned to every word in a corpus in such a way that words with similar contexts (and thus, similar meanings) are placed close to one another within this embedding. Thus, the similarity of two words can later be determined by applying a distance measure to their vectors, that is, by measuring their proximity in the embedding (cf. below). Another useful property of such

9 Zellig S. Harris, "Distributional structure", *Word* 10 (23) (1954), 146–162.

an embedding is that it reflects not only semantic similarity but also semantic relations, as illustrated by the example in figure 1 (the vector space with originally 100 dimensions is here converted with a principal component analysis (PCA) into a two-dimensional one for the sake of presentability): it shows the relations between four Greek adjectives and their antonyms as dotted lines. Note that the lines form a pattern: it indicates that the semantic relation between adjective and antonym is similar for each pair which becomes measurable (and in this case visible) in the vector space.

Fig. 1: A convenient 2d-PCA projection of four pairs of Greek adjectives with their antonyms (from left to right): δύνατος (possible) – ἀδύνατος (impossible), δῆλος (clear) – ἄδηλος (unclear), χρηστός (useful) – ἄχρηστος (useless), μικτός (mixed) – ἄμικτος (unmixed).



We achieved a reliable word embedding of our corpus (the TLG), which means assigning a vector to each word depending on the contexts in which it is used, with the aid of word2vec[10] in combination with a comprehensive semi-automatic evaluation of different parameter sets and normalisations. The principle of measuring the similarity of the meaning of two words by their proximity in the embedding can be

10 Thomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean, "Efficient estimation of word representations in vector space", *Proceedings of the International Conference on Learning Representations in Scottsdale, Arizona* (2013), n. pag.

extended to whole phrases, for example, with the Word Mover's Distance (WMD),[11] which we use for our approach to paraphrase search. By applying such a measure, it is possible to automatically compare two phrases with respect to their meaning. In simplified terms, the WMD matches the individual words of the two phrases so that they form pairs (with the most similar ones being coupled in respect of their distance within the embedding); the average of the distances among these pairs represents the distance between the two phrases: the smaller the distance, the more similar the passages. For instance, the distance between a word and a verbatim citation of it naturally amounts to zero. If the word is used in another form or a synonym is applied instead, the distance is small; and if the word is exchanged for a semantically unrelated expression, the distance is large.

The tool which is optimised to work on the Ancient Greek corpus can be applied to a given passage of Plato in order to search for similar phrases within the corpus. It is also applicable to other research questions which involve searches for similar passages, as for example the search for sources of a given passus instead of adaptations.

Since the WMD is very costly to calculate and millions of calculations have to be performed in order to search the entire TLG, the first version developed by our team was too slow to offer an interactive search where the reader can spontaneously repeat the search and refine the results by adapting, for example, the length of the text section he is interested in or by changing the parameters. To address this problem, we used a simplified but still reliable heuristic, the relaxed Word Mover's Distance (rWMD), to pre-filter reasonable paraphrase-candidates first. In addition, we achieved a drastic reduction in the runtime of the approach by taking advantage of a mathematical special case that allows partial precomputation and also tolerates fuzziness in the length of paraphrase candidates.[12] The latter yielded a significant improvement regarding our research questions since paraphrases and other forms of intertextual relations, especially in literature, often differ in length from the original passage they refer to because they tend to abridge or extend it. This generally poses one of the bigger challenges for automated paraphrase detection which we were able to overcome in this way.

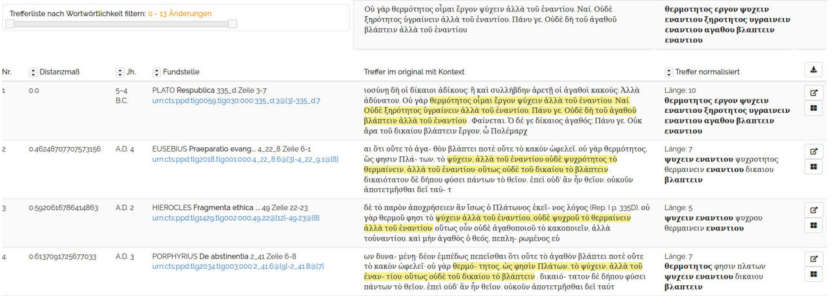
Together with further optimisations, the rWMD-search thus became usable in real time which significantly improved user experience and the way it is applied (cf. below). Figure 2 shows an excerpt of a hit list that is displayed after a search with

11 See Matt J. Kusner, Yu Sun, Nicolas I Kolkin, and Kilian Q. Weinberger, "From word embeddings to document distances", in: Francis Bach and David Blei (ed.): *Proceedings of the 32th International Conference on Machine Learning*, Volume 37 (Lille, France: JMLR.org, 2015), 957–966.

12 See for more details Marcus Pöckelmann, Janis Dähne, Jörg Ritter and Paul Molitor, "Fast paraphrase extraction in Ancient Greek literature", *it – Information Technology* 62, 2 (2020), 75–89.

the rWMD for potential paraphrases to a passus from Plato's *Republic* (335d3–7). In the passus, Plato establishes the influential axiom that heat does not cool and the good does not harm. The passage itself occupies the top of the list (No. 1) as a perfect match with the search terms (the distance amounts to zero). In the second place, we find an adaption by the Christian author Eusebius of Caesarea (*Praeparatio Evangelica* 4.22.8.5–9.1) with many verbatim agreements.

Fig. 2: The first four hits of a larger list of paraphrase candidates generated with the rWMD-search. The hits along with their distance and location are displayed as full text together with some context and also in the normalised form processed by the algorithm. In addition, there is the option to import a hit directly into the Reference-Annotator.



The reflection on the digital methods also had consequences for our view on the literary phenomena: we developed a new perspective on the concepts of *paraphrase* and *intertextuality* in tune with our digital approaches, especially the representation of the corpus in a vector space.[13] Some of our methodological insights were fuelled by the evaluation of the output of our algorithms and, in turn, mainly had the following consequences for the way we designed the representation of the results: paraphrase, as a phenomenon belonging to the field of intertextuality, is intrinsically linked to the recipient.[14] In literature and in other forms of communication, a text segment only becomes a paraphrase if a recipient establishes an intertextual relation between this segment and another (the supposed original phrase); otherwise, the text simply remains a phrase. Applied to literature: if for example James Joyce had designed a passage in his *Ulysses* as an allusion to Homer's *Odyssey* by paraphrasing a passus of

13 Cf. Kurt Sier and Eva Wöckener-Gade, „Paraphrase als Ähnlichkeitsbeziehung. Ein digitaler Zugang zu einem intertextuellen Phänomen", in: Charlotte Schubert, Paul Molitor, Jörg Ritter, Joachim Scharloth, and Kurt Sier (ed.): *Platon Digital: Tradition und Rezeption* (Heidelberg: Propylaeum, 2019), 23–43.

14 Cf. in general the conception of Susane Holthuis, *Intertextualität. Aspekte einer rezeptionsorientierten Konzeption* (Tübingen: Stauffenburg Verlag, 1993).

the epic but no reader had ever linked Joyce's passage back to the Homeric one, intertextuality would not have been established, and Joyce's words would have stayed a phrase of his own design for everybody except for him. Moreover, the paths of the reception of Plato are so entangled and so many waypoints have been lost that one can rarely say with certainty whether a passage directly refers to Plato or rather echoes him indirectly, for example, through an intermediate source. Thus, we chose to consider a paraphrase basically as a phrase with the potential to become a paraphrase whenever an intertextual relation is established by someone, and the author simply as the first one who did so. This recipient-centred conception is also reflected in our approaches to paraphrase search, which we conceptualised as semi-automatic: both the paraphrase searches we developed, the one via n-grams (which we cannot address in depth in this paper) and the other via rWMD, propose a list of text passages which meet certain criteria to rank as promising paraphrase candidates (cf. below). The output can be sorted according to various criteria and thus adjusted to different research questions: either according to the similarity, or going by the dates of the texts' authors so that an assessment is possible from a historical perspective. The results can be exported as csv-files (comma-separated values) for further processing.

Which of the results of the search should be considered as paraphrases wholly depends on the judgement of the individual researcher. Experts might disagree here in many cases (as we have seen in our team and in literature on paraphrasing), thus encouraging a fruitful scholarly debate. For example, a researcher may classify a passus from a later author which resembles an aphorism coined by Plato as a sophistically modified paraphrase adapted directly from his dialogues; another may regard it only as a faint resonance of the passage mediated through the later tradition. Hopefully, they will discuss their divergent assessments and seek further evidence to support their views, thus contributing to the exploration of Plato's reception in antiquity.

Detecting paraphrases: between similarity and vagueness

Concerning the necessary but insufficient characteristics of a paraphrase in relation to the initial phrase (or pre-text), which we want to find automatically, both similarity and, often neglected, vagueness play a central role in their detection:[15] every paraphrase exhibits a significant similarity to the pre-text which enables the recipient to establish the intertextual relation between the two passages in the first place, as outlined above. However, this similarity can appear in many diverse forms and

15 Cf. Manfred Pinkal, *Logik und Lexikon – Die Semantik des Unbestimmten* (Berlin/New York: De Gruyter, 1985).

to different degrees (vagueness)—an almost exact match with the pre-text moves it closer to a citation and is easier to detect whereas loose verbal resemblances are often found in the context of allusions and similar literary phenomena which are more challenging for the readers and their stock of knowledge. Here, the approach based on the Word Mover's Distance yielded very promising results because it works almost without predefined criteria and can detect both stronger and weaker similarities between the passages as indicated by the measured distance.

Detecting paraphrases is like fishing: new methods

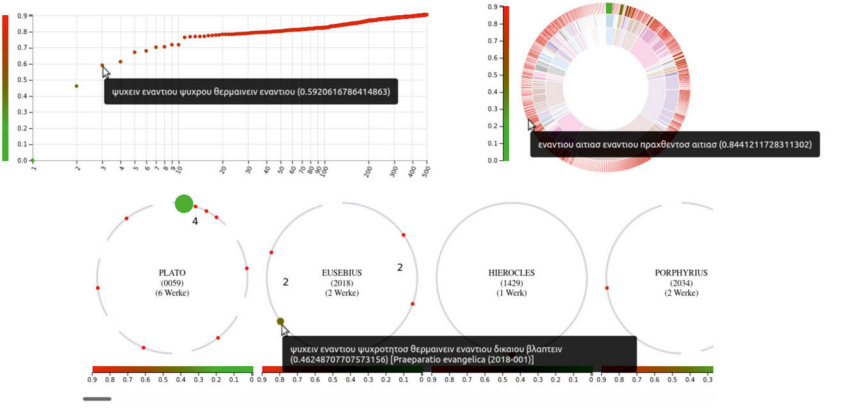
Regarding both of our approaches to paraphrase search, it is important to stress that the results they yield should never be regarded as either complete or exact which is not a weakness in our eyes but potentially advantageous. In the traditional scholarly approach, after singling out a passage or text, one would search in a very focused way for parallels in predefined texts and/or for predefined single words. In contrast, our tools use very few predefinitions. One could compare both methods to fishing with a rod (choosing a promising spot, using a specific hook and bait, devoting a lot of time, etc.) and fishing on a trawler with a net. Using the rod, with a little patience, we will probably catch the special fish we hoped for. Using the net, we catch many different types of fish at the same time: some interesting, some not, but most importantly, many unexpected ones. What was surprising for our team was the difference that the change in the speed of obtaining the results made for the way we chose and adapted our research subjects; it definitely opened new possibilities to learn by trial and error. It was also helpful for evaluating and further improving the paraphrase search and getting to know the best way to apply it since more and different types of searches (e.g. for shorter or longer phrases, using different parameters such as ignore/include stop words) could be explored and compared.

Visualising paraphrases: new perspectives

The increasing number of potential results also poses challenges to the processing of the output. Here, besides sorting the candidates suggested as paraphrases according to various parameters, visualisation comes into play. We implemented various types of visualisation to provide the researcher with a quicker and more intuitive access to the results in order to 'give him an overview'. It is important for us to clarify that also in the visualised form, the output of the paraphrase search is not conceived to represent final results, but to leave room for interpretation by the individual researcher. We tried to counteract the impression of definitiveness by implementing three types of distant reading visualisations at once which yield diverse pictures of

the transformed search results: A scatter plot for illustrating the distribution of the distance values, an interactive and multi-level sunburst diagram as an overview of the involved authors, works, and passages suggested as paraphrase-candidates, and thirdly a *planetary* visualisation as a more detailed but also more space-consuming alternative to the sunburst diagram, see figure 3 (the phrase on which the search is based is again the axiom from Plato's *Republic* 335d3–7 mentioned above).

Fig. 3: The three distant reading visualisations we integrated into our digital environment for illustrating the results of a paraphrase search. Left: a scatter plot for the distance values. Right: a sunburst diagram which aggregates authors and works for all search hits. Bottom: the search hits broken down in detail for individual authors.

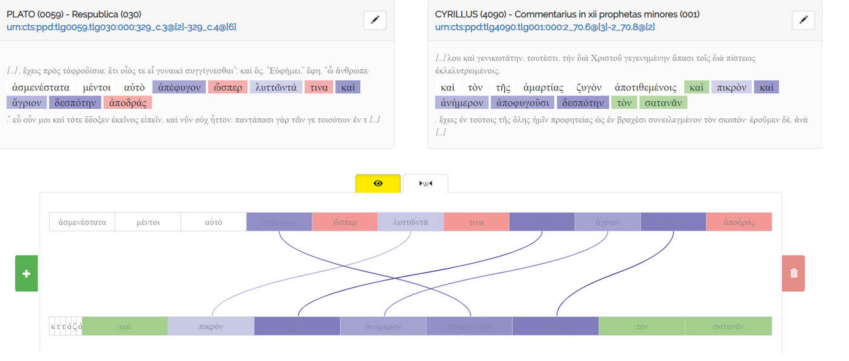


On the other hand, for the Reference-Annotator, a tool developed by us for microanalysis of confirmed paraphrases, we successfully implemented visualisations which represent the relations between a paraphrase and the original phrase in a way that makes them more easily comprehensible and comparable to other instances.[16] One example is shown in figure 4: at the beginning of Plato's *Republic*, an old man relates that he is relieved to have escaped from libido like a runaway slave from a choleric and savage master (329c3–4). The literary simile has been adapted by the Church Father Cyril of Alexandria with the flight from libido being converted into the one from Satan (*Commentary on the Twelve Prophets* 70, 6–8). The Reference-Annotator allows the scholar to identify word-to-word relations between the passages and classify them according to linguistic categories (e.g. 'same lemma, same form' or

16 See Marcus Pöckelmann and Eva Wöckener-Gade, "Bridging the Gap between Plato and his Predecessors. Towards an Annotated Gold Standard of Intertextual References to Plato in Ancient Greek Literature", Presentation at the EADH (2018), 07.-09.12.2018 in Galway.

'same lemma, different form', visualised in different shades of blue). Moreover, additions and omissions of words can be marked (visualised in green/red). This close-reading visualisation of the annotator helps to overcome obstacles in the analysis of paraphrases based on written text alone, which is apparent from the amount of competing linguistic approaches mentioned above.

Fig. 4: A double visualisation of the word-to-word relations between a passage from Plato's Republic and its adaption by Cyril of Alexandria in the Reference-Annotator. Above, the text of Plato is displayed on the left and the paraphrase by Cyril on the right. Different types of relations between the words can be annotated (visualised as lines in various shades of blue) and both additions (green) and omissions (red) can be marked with the tool, which results in the visualisation in the lower part of the image.



Conclusion

The project Digital Plato does illustrate how interdisciplinary teamwork and mixing methods in an iterative process can yield new insights. In our case, starting from our research question and rejecting already existing approaches as not applicable thereto, not only new and innovative digital tools were developed in accordance with our hermeneutical framework but the engagement with their theoretical foundations also led us to rethink the literary conceptions involved. The insights thus gained led to further adaptations of our algorithms and tools, and to the development of new components such as visualisations and the Reference-Annotator as a tool for the analysis of paraphrases. The insights and tools presented in this paper stem from the work of all team members, including numerous research assistants from the differ-

ent disciplines, as illustrated by our project monograph,[17] who deserve our special thanks.

Bibliography

Bhagat, Rahul, and Eduard Hovy. "What Is a Paraphrase?". *Computational Linguistics* 39 (2013): 463–472. DOI: https://doi.org/10.1162/COLI_a_00166.

Blackwell, Christopher, and David Neel Smith, "The Canonical Text Services protocol". GitHub cite-architecture. Last modified Apr 10, 2014. https://github.com/cite-architecture/cts_spec/blob/master/md/specification.md

De Beaugrande, Robert-Alain and Wolfgang U. Dressler. *Einführung in die Textlinguistik* (Konzepte der Sprach- und Literaturwissenschaft 28). Tübingen: De Gruyter, 1981.

Harris, Zellig S. „Distributional structure". *Word* 10 (23) (1954): 146–162.

Ho, Chuk F., Masrah Azrifah Azmi Murad, Shyamala Doraisamy, and Rabiah A. Kadir. "Extracting lexical and phrasal paraphrases: a review of the literature". *Artificial Intelligence Review*, 42.4 (2012) : 851–894. DOI : https ://doi.org/10.1007/s10462-012-9357-8.

Holthuis, Susanne. *Intertextualität. Aspekte einer rezeptionsorientierten Konzeption*. Tübingen: Stauffenburg Verlag, 1993.

Kristeva, Julia. "Wort, Dialog und Roman bei Bachtin (1967)". In *Literaturwissenschaft und Linguistik. Ergebnisse und Perspektiven. Band 3: Zur linguistischen Basis der Literaturwissenschaft II*, edited by Jens Ihwe, 345–375. Frankfurt am Main: Athenäum, 1972.

Kusner, Matt J., Yu Sun, Nicholas Kolkin, and Kilian Weinberger. "From word embeddings to document distances". In *Proceedings of the 32th International Conference on Machine Learning* 37, edited by Francis Bach and David Blei, 957–966. Lille, France: JMLR.org, 2015.

Mikolov, Thomas, Kai Chen, Greg Corrado, and Jeffrey Dean. "Efficient estimation of word representations in vector space". In *Proceedings of the International Conference on Learning Representationsin Scottsdale, Arizona* (2013): n. pag.

Pfister, Manfred. "Konzepte der Intertextualität". In *Intertextualität: Formen, Funktionen, anglistische Fallstudien* edited by Ulrich Broich and Manfred Pfister, 1–30. Tübingen: De Gruyter, 1985.

Pinkal, Manfred. *Logik und Lexikon – Die Semantik des Unbestimmten*. Berlin/New York: De Gruyter, 1985.

Pöckelmann, Marcus and Eva Wöckener-Gade. „Bridging the Gap between Plato and his Predecessors. Towards an Annotated Gold Standard of Intertextual Refer-

17 Charlotte Schubert, Paul Molitor, Jörg Ritter, Joachim Scharloth, and Kurt Sier (eds.). *Platon Digital: Tradition und Rezeption* (Heidelberg: Propylaeum, 2019).

ences to Plato in Ancient Greek Literature". Presentation at the EADH (2018), 07.-09.12.2018 in Galway. Abstract: https://eadh2018.exordo.com/files/papers/49/final_draft/Poeckelmann_Woeckener-Gade_Bridging_the_gap.pdf

Pöckelmann, Marcus, Janis Dähne, Jörg Ritter, and Paul Molitor. "Fast paraphrase extraction in Ancient Greek literature". *it – Information Technology* 62, 2 (2020): 75–89. DOI: https://doi org/10.1515/itit-2019-0042.

Schubert, Charlotte, Paul Molitor, Jörg Ritter, Joachim Scharloth, and Kurt Sier (eds.). *Platon Digital: Tradition und Rezeption* (Digital Classics Books, Band 3), Heidelberg: Propylaeum, 2019. https://doi.org/10.11588/propylaeum.451

Sier, Kurt and Wöckener-Gade, Eva: 'Paraphrase als Ähnlichkeitsbeziehung. Ein digitaler Zugang zu einem intertextuellen Phänomen'. In: *Platon Digital: Tradition und Rezeption*, edited by Charlotte Schubert, Paul Molitor, Jörg Ritter, Joachim Scharloth, and Kurt Sier, 23–43. Heidelberg: Propylaeum, 2019. DOI: https://doi.org/10.11588/propylaeum.451

Vila, Marta, M. Antonia Martí, and Horacio Rodríguez. "Is This a Paraphrase? What Kind? Paraphrase Boundaries and Typology". *Open Journal of Modern Linguistics* 4 (2014): 205–218. DOI: https://doi.org/10.4236/ojml.2014.41016

Wilkinson, Mark D., Michel Dumontier, IJsbrand J. Aalbersberg et al. 'The FAIR Guiding Principles for scientific data management and stewardship'. *Sci Data* 3, 160018 (2016): n. pag. DOI: https://doi.org/10.1038/sdata.2016.18

#SIMILARITY

In the Oxford English Dictionary (OED), the term #Similarity initially describes an everyday meaning of likeness or resemblance and the ensuing details or aspects "in which a particular thing is similar to another". Further uses and compounds largely focus on "systems or processes for indicating or determining the degree of similarity between particular things" with a strong reference to mathematics. *

From this, the mixed method projects derive the digital humanities uses of the term, in which common comparisons play a minor role. The focus seems to be on identifying degrees of similarity and relating them to each other or mapping them in absolute terms. "Similarity is a relation between two entities, which is characterized by their significant proximity to a tertium comparationis" (Digital Plato). The entities vary and may include for example numerical, semantic or stylistic properties. As such, the concept of similarity may shift into an entirely abstract digital realm accessed by #Modelling and #Human-in-the-loop practices: "Among the steps that lead from word frequencies to a quantification of stylometric similarity are lemmatization, filtering, occurrences and dispersion, normalization, and the use of different aggregation functions" (ReadingAtScale).

The findings are often interpreted and discussed by clustering and/or #Visualization, thus connecting back to the realms of humanities' practices of comparison (ReadingAtScale, ANCI).

* "similarity, n.". in: *Oxford English Dictionary* (OED), Third Edition, September 2019; most recently modified version published online March 2022, https://www.oed.com/ [accessed: 20.05.2022].

Title: Reading at Scale
Team: Thomas Weitin, Simon Päpcke, Katharina Herget, Anastasia Glawion, Ulrik Brandes
Corpus: Deutscher Novellenschatz (Heise & Kurz, Eds., 1871–1876)
Field of Study: German Literature, Statistics, Network Science
Institution: TU Darmstadt, ETH Zürich
Methods: Stylometry, Network Analysis, Text Analysis
Tools: Natural Language Processing, Statistics, Clustering, Questionnaire
Technology: R, visone, own implementations

Reading at Scale. A Digital Analysis of German Novellas from the 19th Century (Reading at Scale)

Thomas Weitin, Simon Päpcke, Katharina Herget, Anastasia Glawion, Ulrik Brandes

> "das Beste, was in dieser Gattung geleistet ist […] zu sammeln und in übersichtlicher Folge herauszugeben, bedarf wohl kaum der Rechtfertigung".[1]

Abstract *The Deutscher Novellenschatz (published in 24 volumes 1871–1876) is a collection of 86 German-language novellas edited by Paul Heyse and Hermann Kurz. It is an example of a medium-sized corpus amenable, in principle, to both, scholarly reading and automated analysis. Our point of departure was the conviction that research questions at intermediate granularity would require the combination of hermeneutic and statistical methods to achieve an appropriate level of abstraction while maintaining a sufficient amount of context. By exposing the sensitivity of text similarity measures to choices in the preparation and evaluation of bag-of-word representations, we highlight the need for consideration of contextual information even in the most distant reading approaches. Literary theory suggested more coarse-grained hypotheses that we tested both empirically in a group of non-expert readers and computationally using similarity of character constellations, respectively. Based on correspondences of the editors and comparison with other corpora, the Novellenschatz was further situated in a historiographic context.*

Introduction

An important yet rarely studied issue in Digital Philology as a subdomain of the Digital Humanities is the composition of text corpora. Despite the availability of vari-

1 "To collect the best that has been achieved in this genre […] and to publish it in a well arranged order hardly needs justification", translated by the authors. Paul Heyse and Hermann Kurz, "Einleitung", in *Deutscher Novellenschatz*, vol. 1 (München: R. Oldenbourg, 1871).

ous repositories,[2] there is no gold standard for corpus composition in digital literary studies. Such a standard is perhaps unattainable, because literary text corpora must be prepared individually and purposefully according to the specific research question of the project at hand. Even more so, the constitution of digital literary corpora has to be critically reflected upon.

Within the 'Reading at Scale' project, we focused on the novella collection *Deutscher Novellenschatz*. One of the distinctive features of this research object is that it is both a corpus and an artefact at the same time: With its overall 213 novellas, the *Novellenschatz*-series—including the *Deutscher Novellenschatz* (DNS), the *Neuer Deutscher Novellenschatz* (NDNS) and the *Novellenschatz des Auslandes* (NSdA)[3]—comprises a number of popular German-language novellas (and in the case of the *NSdA* foreign-language novellas translated into German) suitable for statistical analysis.[4] As an artefact, the *Novellenschatz*-collection is an object of historical interest, specifically with regard to its normative claims showcased in the quotation above. The novella collection is complemented by letters that the main editor Paul Heyse exchanged with the others: Hermann Kurz until his passing in 1873 and Ludwig Laistner thereafter. In the letters, the editors thoroughly discuss some of the choices they made for the collection. This correspondence, which is unfortunately only published in excerpts,[5] highlights that the notion of relationality was of particular relevance for the selection process.

The so-called long 19th century (1789–1914) is well known for its rapidly growing mass market for literature, including magazines and newspapers.[6] In addition, the novella—especially the Realistic novella—became the epitome of this mass production in the German-speaking literary world. This progressive development cul-

2 For example: "Deutsches Textarchiv. Grundlage für ein Referenzkorpus der neuhochdeutschen Sprache. Herausgegeben von der Berlin-Brandenburgischen Akademie der Wissenschaften", http://www.deutschestextarchiv.de/, or *Textgrid*, TextGrid Consortium. 2006–2014. TextGrid: A Virtual Research Environment for the Humanities. Göttingen: TextGrid Consortium. textgrid.de.

3 Paul Heyse and Hermann Kurz (ed.), *Deutscher Novellenschatz* (1871–1876, Bd. 1–24, 86 novellas) München: R. Oldenbourg. Paul Heyse and Ludwig Laistner (ed.), *Neuer deutscher Novellenschatz* (1884–1887, Bd. 1–24, 70 novellas) München: R. Oldenbourg. Paul Heyse and Hermann Kurz (ed.), *Novellenschatz des Auslandes* (1877–1884, Bd. 1–14, 57 novellas) München: R. Oldenbourg.

4 The collection is completely digitized and is already being published (Weitin 2016; 2018).

5 Monika Walkhoff, *Der Briefwechsel Zwischen Paul Heyse Und Hermann Kurz in Den Jahren 1869 – 1873 Aus Anlass Der Herausgabe Des Deutschen Novellenschatzes [Mit Faks.]* (München: Foto-Druck Frank, 1967).

6 Matt Erlin and Lynne Tatlock, "Introduction: 'Distant Reading' and the Historiography of Nineteenth Century German Literature", in *Distant Readings: Topologies of German Culture in the Long Nineteenth Century*, ed. Matt Erlin and Lynne Tatlock, Studies in German Literature, Linguistics, and Culture (Rochester, New York: Camden House, 2014), 1–29.

minates in the *Novellenschatz*-approach "to collect the best that is produced in this genre […] and to publish it in a clear sequence".[7] By setting this requirement, the collection not only reflects the contemporary perception of a literary deluge ('Literaturschwemme') but also responds to it with the programmatic preface that includes the falcon theory ('Falkentheorie'), an instrument to assess the quality of Realistic novellas[8] which led to high esteem for Heyse as a theorist of the novella genre accorded by literary historian Oskar Walzel.[9]

The requirement proposed by the editors mirrors the standards of literary quality established by 19th-century literary historiography, whereby the quality of the individual text is defined in relation to the epoch, its ideals and predecessors.[10] Heyse and Kurz demonstrate it by opening the *Novellenschatz*-series with Goethe's fairytale *Die Neue Melusine* to promote their underlying Realistic agenda in contrast to the classic example of Romanticism.[11] Heyse contextualizes this choice as follows:

> "In these first volumes we must carefully avoid offending the big bunch and save our caviar for the middle of the table, where they have already learned to swallow all kinds of things."[12]

The quote demonstrates that the selection of novellas, their order and position within the collection are strongly affected by their relation to other texts. This is not a notion that is distinctive for the *Novellenschatz*, but rather a constant underlying feature of the literary market that writers of the 'Literaturschwemme' had to take into account, as they were writing with the awareness of broad similarity.[13] Consid-

7 Paul Heyse and Hermann Kurz, "Einleitung", in *Deutscher Novellenschatz*, vol. 1 (München: R. Oldenbourg, 1871), https://www.deutschestextarchiv.de/heysekurz_einleitung_1871, translated by the authors.

8 See Thomas Weitin and Katharina Herget, "Falkentopics: Über Einige Probleme Beim Topic Modeling Literarischer Texte", *Zeitschrift Für Literaturwissenschaft Und Linguistik* 47, no. 1 (2017): 29–48, https://doi.org/10.1007/s41244-017-0049-3.

9 Wilhelm Scherer and Oskar Walzel, *Geschichte Der Deutschen Literatur* (Berlin: Askanischer Verlag, 1921).

10 E.g. Oskar Walzel, *Die Deutsche Dichtung Seit Goethes Tod* (Berlin: Askanischer Verlag, 1919); Wilhelm Scherer, *Geschichte Der Deutschen Litteratur*, 3. Auflage (Berlin: Weidmannsche Buchhandlung, 1885); Georg Gottfried Gervinus, *Handbuch Der Geschichte Der Poetischen National-Literatur Der Deutschen*, 3. Aufl. (Leipzig: Engelmann, 1844).

11 See Thomas Weitin, *Digitale Literaturgeschichte. Eine Versuchsreihe in 7 Experimenten* (Berlin: Metzler/Springer Nature, 2021).

12 "Wir müssen gerade in diesen ersten Bänden sorgfältig vermeiden dem großen Haufen vor den Kopf zu stoßen und unsern Caviar lieber für die Mitte der Tafel sparen, wo sie schon allerlei schlucken gelernt haben". Unpublished correspondence letter, P. Heyse to H. Kurz, 14.10.1870, translated by the authors.

13 Thomas Weitin, "Average and Distinction. The Deutsche Novellenschatz Between Literary History and Corpus Analysis", LitLab Pamphlet, 6 (2018): 1–23.

ering the influence of text relationality on novella production and on the selection processes for the collections in question, we were particularly interested in methods that reflected these connections. Thus, this publication first presents our thoughts on issues of comparability within a literary corpus in general, and in the second part delves into different preprocessing techniques that affect groupings within the first *Novellenschatz*-collection.

Between contextualization and commensuration

When Franco Moretti coined the term 'distant reading' 20 years ago, he certainly did not foresee its subsequent evolution. Originally, Moretti used it to promote his idea to discover the great unread of world literature: in opposition to close reading, distant reading should transgress the limits of the Western canon by using synthesis/meta-analysis "without a single direct textual reading".[14] Regardless of the original meaning, the term was quickly adopted by digital literary scholars. Although Moretti later admitted that the term was originally meant as a joke,[15] its reception certainly influenced the foundation of the Stanford Literary Lab in 2010. Building upon this development, Martin Mueller introduced the term 'scalable reading' as a "happy synthesis of 'close' and 'distant' reading",[16]which allows one to "zoom" between single text and corpus level through digital surrogates.[17]

As a result of our own corpus analyses, we find the scaling metaphor to be misleading. While we do share Mueller's assertion that the "typical encounter with a text is through a surrogate",[18] no matter whether this means reading a single book edition or analysing the data of an entire corpus, we feel that it is limiting to assume the existence of dimensions along which surrogates can be transformed into one another via scaling. From our interdisciplinary experience as professional readers of literature and analysts of data, it is neither enough to state that "the two scales of analysis […] need to coexist",[19] nor is it helpful or even necessary to expect the different surrogates to correspond to the levels of detail in the same representation.

---

14 Franco Moretti, "Conjectures on World Literature", *New Left Review* 1 (2000): 57.

15 Franco Moretti, "Conjectures on World Literature", in *Distant Reading* (London; New York: Verso, 2013), 43–62.

16 Martin Mueller, *Scalable Reading*, 2019, https://scalablereading.northwestern.edu/.

17 Thomas Weitin, "Thinking Slowly. Reading Literature in the Aftermath of Big Data", *LitLab Pamphlet* #1 (2015): 1–19.

18 Martin Mueller, "Morgensternś Spectacles or the Importance of Not-Reading", Northwestern University Digital Humanities Laboratory (blog), 2013, https://sites.northwestern.edu/nudhl/?p=433.

19 Matthew L. Jockers, *Macroanalysis: Digital Methods and Literary History*, Topics in the Digital Humanities (Urbana, IL: University of Illinois Press, 2013), 9.

Instead, we posit that appropriate scales and representations are determined by research question, research design, and project resources. In addition to reading from up close or at a distance, it may be useful to have a variety of angles and lenses at one's disposal. In other words, different representations, or surrogates, may be required even at the same level of abstraction, especially because adjacent levels are generally related in ways that are more complex than simply a change of resolution.

With reading at scale, the methodological problems at stake are the opposing actions of contextualization and commensuration. To understand the subtleties of a case, context helps to distinguish it from others. Some aspects of literature, however, become comparable only if specific features are selected in which they can be treated as being of the same kind, although maybe in different ways.[20] In annotation projects and workflows such as CATMA, this logic is applied to the level of text passages within a singular text.[21] The balanced consideration of discriminating peculiarities and shared features governs what makes an adequate representation and ultimately determines the range of methods available. The specific interpretation of the term contextualization that is relevant here will become more apparent after we relate commensurability to statistical analysis.

Any form of statistical data analysis relies on variables or the assignment of values in a specified range to each entity from a defined domain in order to express levels or qualities of a property they all have in common. Variables are productive only if there is at least the potential for the entities to share the property present in them. A categorical variable encoding the author of a text or a numerical one encoding the number of occurrences of the word 'time' is potentially informative, but a binary variable encoding whether a text starts with a specific phrase is generally not unless the phrase is 'Once upon a time' and the corpus contains fairy tales. For any given text, references to biographic information or past works of an author are highly contextual and rarely suited for variable-based comparison across a corpus.

An example of a variable commonly defined for a literary corpus such as ours is the original publication date of each text. Innocent as this may seem, using such a variable, for example to report the distribution of texts over time, begets a number of assumptions. By referring to the elements of a corpus indiscriminately as texts, we are already suggesting that they are entities of the same kind, or are commensurate. In some sense they are so, because all of them similarly occupy a number of pages in

20 Bettina Heintz, "Numerische Differenz. Überlegungen zu einer Soziologie des (quantitativen) Vergleichs / Numerical Difference. Toward a Sociology of (Quantitative) Comparisons", *Zeitschrift für Soziologie* 39, no. 3 (1 January 2010), https://doi.org/10.1515/zfsoz-2010-0301.

21 Evelyn Gius et al., "CATMA 6", 6 April 2022, https://doi.org/10.5281/ZENODO.1470118; Jan Horstmann, "Undogmatic Literary Annotation with CATMA. Functions, Differentiation, Systematization", in *Annotations in Scholarly Editions and Research*, ed. Julia Nantke and Frederik Schlupkothen (De Gruyter, 2020), 157–76, https://doi.org/10.1515/9783110689112-008.

one of the volumes of the *Novellenschatz*. For a different purpose, considering titles with different text lengths, such as Wolf's *Stern der Schönheit* (around 20,000 characters) and Auerbach's *Diethelm von Buchenberg* (400,000 characters), as similar entities may be questionable. And even when it appears appropriate to treat them as being of comparable type, it is a separate, substantive question whether their publication dates constitute a feature by which they can be compared. The differences between periods of writing, contemporaneous developments, versions of a text, or forms of publication may be too stark to allow a meaningful ordering of texts by publication date, let alone an interpretation of the length of time intervals in-between.

Variables such as publication date, genre, or gender of the author are extrinsic to the text and often referred to as metadata.[22] Technically, metadata are data about data, which implies that texts are considered data themselves. Intrinsic, or text-immanent, variables may associate each text in a corpus with, for instance, a word-frequency vector, linguistic indices, or plot complexity. They are generally based on an intermediate representation that is defined for each text individually, and then summarized into corpus-level variables. A prominent example are character constellations, which represent relationships such as co-occurrences in a scene between characters that are specific to a text.[23] The network variables denoting the co-occurrence of a pair of characters in a scene are different for each text. Since their dramatis personae are different, they do not assign values to the same pairs of characters, but represent each text in its own specific way. Characteristics such as an index of centralization for co-occurrence networks, however, are comparable across texts and therefore similar to other text descriptors.

Intermediate representations such as character constellations focus on particular aspects of a text and, therefore, filter out details and abstract more general features from the specificities of a text, but they do so for each text individually. Comparability arises from the shared structure of these text-specific representations. The crucial decision thus lies in the level of detail that, on the one hand, needs to be preserved to sustain important distinctions between texts and that, on the other, needs to be abstracted from to allow comparisons across a corpus. There is, however, no inherent relation between intermediate representations, or partial abstractions, of texts that would correspond to a notion of scaling. Bag-of-words representations, character networks, and event sequences surely represent a text on different levels

22 Matteo Lorenzini, Marco Rospocher, and Sara Tonelli, "On Assessing Metadata Completeness in Digital Cultural Heritage Repositories", *Digital Scholarship in the Humanities* 36, no. Supplement 2 (5 November 2021): 182–88, https://doi.org/10.1093/llc/fqab036; Christof Schöch, "Big? Smart? Clean? Messy? Data in the Humanities", *Journal of Digital Humanities* 2 (2013): 1–12.

23 See Benjamin Krautter et al., "Eponymous Heroes and Protagonists – Character Classification in German-Language Dramas", *LitLab Pamphlet* #7 (2018): 58, https://www.digitalhumanitiescooperation.de/en/pamphlete/pamphlet-7-interpretierbare-figurenklassifikation/.

of abstraction, but neither of these levels refines or coarsens another. The appropriate choice of scale is determined substantively by the research question, but also pragmatically by the possibility for objective and manageable realization.

Reading at scale is thus fundamentally about identifying a suitable combination of qualitative and quantitative aspects to determine representations, or mixing methods. In data analysis, the qualitative and quantitative notions often refer to the level of measurement of a variable. Nominal and ordinal variables have values representing (un)ordered categories. They do not admit arithmetical operations and are therefore considered qualitative data, whereas values on a ratio-scale express multiples of a unit and are therefore quite literally quantitative. Consequently, there can be qualitative and quantitative variables, and the distinction is made based on properties of the range of values. Since literary epochs generally do not progress linearly, one might argue that publication dates are on an ordinal level of measurement at most.

In hermeneutics, the discomfort with quantitative analyses appears to occur even earlier: often, the very assumption that information can be represented meaningfully in variables is resented. In other words, the commensurability of entities is questioned because every such attempt would rid the subject of crucial circumstantial evidence.

Every mixed-method approach faces this kind of trade-off, and we can work from both ends to arrive at a scale.[24] If we decide to compare the single texts of our corpus only through vectors of word frequencies, the bag-of-words model can be enriched with metadata so that the analysis of historical context becomes a question of subsetting. In Natural Language Processing, word embeddings can be used to foster context sensitivity with respect to both semantics and syntax.[25] And even the results of stochastic semantics as in topic models can be recontextualized through concordance analyses.[26] On the other side, the century-long history of modern philology provides us with resilient methods to avoid drowning in the contextual associations

24 Rabea Kleymann, "Datendiffraktion: Von Mixed zu Entangled Methods in den Digital Humanities", HTML,XML,PDF, ed. Manuel Burghardt et al., *Fabrikation von Erkenntnis – Experimente in den Digital Humanities (Zeitschrift für digitale Geisteswissenschaften / Sonderband)*, 2022, https://doi.org/10.17175/SB005_008; Andrew Piper, *Enumerations: Data and Literary Study* (Chicago, IL; London: The University of Chicago Press, 2018).

25 See Simon Hengchen et al., "A Data-Driven Approach to Studying Changing Vocabularies in Historical Newspaper Collections", *Digital Scholarship in the Humanities* 36, no. Supplement 2 (5 November 2021): 109–26, https://doi.org/10.1093/llc/fqab032; Sahar Ghannay et al., "Word Embedding Evaluation and Combination", in *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)* (LREC 2016, Portorož, Slovenia: European Language Resources Association (ELRA), 2016), 300–305, https://aclanthology.org/L16-1046.

26 See VinhTuan Thai and Siegfried Handschuh, "Context Stamp: A Topic-Based Content Abstraction for Visual Concordance Analysis", in *Proceedings of the 2011 Annual Conference Extended Abstracts on Human Factors in Computing Systems – CHI EA '11* (the 2011 annual confer-

of close reading. For every author, research proves some key passages to be more relevant than others. Depending on new developments, habits, and theoretical umbrella terms in the humanities, neglected parts of literary works become the focus of new readings and interpretations. Within that framework of innovation, a close reading as the dissociative reading of excerpts ('Stellenlektüre') works only as much as a reduction of possible contexts as the apparatus of secondary literature does, to which scholars turn first when they start to explore new research areas.

Stylometric text similarities revisited

As a concrete example on the highest level of context reductions, we next discuss stylometric similarity in the context of corpus analysis. This section draws heavily on our recent study within the scope of the *Novellenschatz*,[27] where many more aspects and detailed examples can be found.

It is notable that even at this almost extreme end of distant reading, many opportunities exist for contextualization, but they are in fact seldom realized due to the failure to adapt to the specifics of a corpus.[28] In bag-of-words representations, the surrogates into which text are abstracted are vectors that assign numerical values to each word in a list deemed relevant for the texts. Such quantification of literary texts naturally comes with the risk of arbitrariness and misinterpretation of artefacts produced by the approach itself.[29] A detailed understanding of every step in the process of operationalization is required to draw defensible conclusions. Much research on stylometric corpus analyses has focused on finding an adequate text distance measure as a means of authorship attribution.[30] In a deliberately compiled corpus, as opposed to a representatively sampled one, we would presuppose the existence of

---

ence extended abstracts, Vancouver, BC, Canada: ACM Press, 2011), 2269, https://doi.org/10.1145/1979742.1979906.

27 Simon Päpcke et al., "Stylometric Similarity in Literary Corpora: Non-Authorship Clustering and 'Deutscher Novellenschatz'", *Digital Scholarship in the Humanities*, 38, no. 1 (2023): 277–95, 1–19, https://doi.org/10.1093/llc/fqac039.

28 Even if there are interesting approaches, e.g. Laura Rettig, Regula Hanggli, and Philippe Cudre-Mauroux, "The Best of Both Worlds: Context-Powered Word Embedding Combinations for Longitudinal Text Analysis", in *2020 IEEE International Conference on Big Data (Big Data)* (2020 IEEE International Conference on Big Data, Atlanta, GA, USA: IEEE, 2020), 4741–50, https://doi.org/10.1109/BigData50022.2020.9377955.

29 Andrew Piper, *Can We Be Wrong? The Problem of Textual Evidence in a Time of Data*, Cambridge Elements. Digital Literary Studies (Cambridge: Cambridge University Press, 2020), https://doi.org/10.1017/9781108922036.

30 John Burrows, "'Delta': A Measure of Stylistic Difference and a Guide to Likely Authorship", *Literary and Linguistic Computing* 17, no. 3 (1 September 2002): 267–87, https://doi.org/10.1093/llc/17.3.267; J. Rybicki, D. Hoover, and M. Kestemont, "Collaborative Authorship: Conrad,

meaningful groups of texts based on attributes other than authorship. Examples include genre, author gender, epoch, or narrative aspects; some of which may be reflected stylometrically. Indeed, even in the seemingly confined space of stylometric operationalizations of similarity, there are sufficient degrees of freedom to tailor them to the different kinds of groups.

Among the steps that lead from word frequencies to quantification of stylometric similarity are lemmatization, filtering, occurrences as well as dispersion, normalization, and the use of different aggregation functions. The resulting similarities are then interpreted through clustering and visualization.

Initial steps include the decision as to which lexical items are retained for further analysis. From a methodological point of view, the selection of terms can rule out undesired associations arising through artefacts from the chosen list of words. Here, the qualification as undesired should always be understood as relative to the underlying research question. Typical differences arising at this stage include decisions on case sensitivity, stemming, and lemmatization. As they are often considered explicitly,[31] we focus on two examples that are different in nature.

Table 1: First-person and other narrative perspectives.

relative frequency of ich vs. first-person narrative	first-person perspective	other perspective
above threshold	16	5
below threshold	5	60

Lists of frequent terms from the documents of a literary corpus usually contain a variety of personal and possessive pronouns. Thus, when we are applying stylometric similarities based on such a list, we can expect to find high degrees of such similarity between texts that are first-person narratives. Table 1 shows that there is a relative-frequency threshold for the first-person singular pronoun 'ich' that serves as a highly accurate classifier for first-person narratives. If, however, the narrative perspective

Ford and Rolling Delta", *Literary and Linguistic Computing* 29, no. 3 (1 September 2014): 422–31, https://doi.org/10.1093/llc/fqu016.

31 E.g. Christopher D. Manning et al., "The Stanford CoreNLP Natural Language Processing Toolkit", in *Association for Computational Linguistics (ACL) System Demonstrations*, 2014, 55–60, http://www.aclweb.org/anthology/P/P14/P14-5010; Nils Reiter, Anette Frank, and Oliver Hellwig, "An NLP-Based Cross-Document Approach to Narrative Structure Discovery", *Literary and Linguistic Computing* 29, no. 4 (January 2014): 583–605, https://doi.org/10.1093/llc/fqu055.

is not the focus of the analysis, it seems to be more fruitful to replace pronouns in the list of frequent words by more generic terms.

Similar effects arise for names of places, countries, cities, or landscapes. While mentions of specific places are usually rare across an entire corpus, they often serve interchangeable roles in their respective novellas. Grouping them into generic tokens such as country, city, or countryside will alter the similarity of texts compared through these terms. A research question could then be guided by epoch theory and test, for example, the hypothesis that texts from Romanticism are placed in rural settings while those from German Realism have urban settings.[32]

The determination of the terms to be included in the analysis establishes the domain of frequency variables, but does not yet specify which values are to be assigned to them. Common choices include the raw counts of words and normalized values based on these counts. However, raw counts do not take into account the distribution of terms across a document. To be able to detect distributional differences, variables may be split into multiple counts related to the same term in different segments of the text. In the context of novellas, segmentation could, for example, be based on the common structuring of the stories within a story. In practice, this is done by comparing texts that contain words with similar document frequency: if we observe a strong accumulation of these terms in a few segments for some texts while others have an even distribution of the terms across the whole document, it suggests that the former texts indeed contain stories within a story.

The options named above are but examples of steps in the data preparation with a potential influence on how texts are related to one another. In order to actually relate the vectors of word frequencies, it is generally advisable to reduce their dimensionality. Otherwise, the word vectors would either have uneven lengths or a high number of zero elements, because many terms occur in only a few of the texts in a corpus. The effect of dimensionality of word vectors on text clustering was, for example, studied by Büttner.[33]

There are three typical filtering methods to avoid this phenomenon and they can also be applied independently. The first is dimensionality reduction by considering only a fixed number of the most frequent words. The second is the use of a stop word list containing common words (usually function words) that are present in each text and thus deemed irrelevant for the distinction of texts. A third possibility, referred to as culling, considers the number of documents in which a term appears: a term is considered only if it appears in a specified percentage of the texts in a corpus. This usually ensures that named entities are not part of the comparison; strong culling

32 See Franco Moretti, *Graphs, Maps, Trees: Abstract Models for a Literary History* (London/New York: Verso, 2005).

33 Andreas Büttner et al., "»Delta«in der stilometrischen Autorschaftsattribution", *Zeitschrift für digitale Geisteswissenschaften*, 2017, https://doi.org/10.17175/2017_006.

can, however, also lead to undesired side effects. In the *Novellenschatz*-corpus, we find that stylometric similarity discriminates the subgenre of Adelsnovellen if noble titles are part of the vector representation, but these titles are generally eliminated by the culling.

As a final example for the choice of representation on a distant reading scale, we consider the values assigned in word frequency vectors. Again there are three major options, the selections of which influence any subsequent analysis. A simple solution is the division of the raw frequency count by text length to make frequencies comparable in different documents. It yields the so-called term-frequency, or tf, score, which amplifies the influence of frequent words on the analysis compared to less frequent words. It can be modified by taking into account the number of documents in which a term appears, thus resulting in the term-frequency inverse document-frequency, or tf-idf, score.[34] This score favours words, which discriminate a text from the majority of the corpus, by attributing higher weights to them. A corresponding finding for the *Novellenschatz* is that novellas of female authors tend to cluster if term-frequency scores are used, owing to their greater use of female pronouns and articles. A scoring method, often used in authorship attribution tasks, is the standardization of term frequencies by subtracting the corpus mean and normalizing this difference by the standard deviation. This so-called z-score expresses higher- and lower-than-average frequencies in units of standard deviation.[35] For the *Novellenschatz*, z-scores prove useful to identify a group of novellas with an unusually high use of verbs in past tense and under-representation of words connected to direct speech. This hints at a group of texts with low prevalence of direct speech, which is deemed uncharacteristic for novellas, as the genre has been claimed to be similar to that of drama.[36]

We have thus argued that, even at a relatively fixed scale, any representation results from a long list of choices. The choices made have consequences and should therefore be made to align with the research question at hand. But it does not end there: when surrogates are processed further, the means of analysis need to be consistent with the goal as well, and not only on a principled level. In the present exam-

34 See Lukáš Havrlant and Vladik Kreinovich, "A Simple Probabilistic Explanation of Term Frequency-Inverse Document Frequency (Tf-Idf) Heuristic (and Variations Motivated by This Explanation)", *International Journal of General Systems* 46, no. 1 (2017): 27–36, https://doi.org/10.1080/03081079.2017.1291635.

35 Thomas Weitin, "Burrows's Delta Und Z-Score-Differenz Im Netzwerkvergleich. Analysen Zum Deutschen. Novellenschatz von Paul Heyse Und Hermann Kurz (1871–1876)", in Fotis Jannidis, Germanistische Symposien (ed.), *Digitale Literaturwissenschaft.* Beiträge Des DFG-Symposiums 2017 (Stuttgart: Metzler, 2023) (forthcoming).

36 "die heutige Novelle ist die Schwester des Dramas", Theodor Storm, "Eine zurückgezogene Vorrede aus dem Jahre 1881", in Albert Köster (ed.), *Theodor Storms sämtliche Werke in acht Bänden*, vol. 8, 2012, 122–23.

ple, vectors of word frequencies, however constructed, are compared pairwise by similarity. Incidentally, the selection of an appropriate measure of similarity is no less important than the construction of the vectors themselves. In fact, the *Novellenschatz* contains examples of three novellas that can be ranked in any order in terms of pairwise similarities by choosing one of the three most common measures. This is mostly due to the different importance these measures assign to large deviations in the scores of single entries.

Conclusion

The reading-at-scale approach critically reflects the trade-off that arises in computational literary analysis: context is reduced to enforce comparability, or comparability is given up for contextualization. Different reduction techniques incorporate a variety of scales for comparison, allowing to view the object of study from different perspectives.

In this chapter, we have demonstrated how different text characteristics affect the values obtained from text distance measures: as the latter are most often computed by averaging distances between words, several signals may be moving and transforming through stages of the analysis. The standard steps for the preprocessing of literary texts for digital analyses, such as tokenization,[37] aim at making texts more comparable and do not focus on eliminating or controlling specific signals. To be able to do that, operationalizations must be reflected upon meticulously and chosen in awareness of potential outcomes that would be considered as artefacts of a method.

With careful consideration, the detection of signals relevant to literary scholarship may be supported by the appropriate choice of representations and parametrization of methods. Beyond existing simplistic suggestions that certain most frequent words harbour the authorship signal, while less frequent words contain the signal for a literary epoch, we have found a variety of other more complex signals in the *Novellenschatz*-corpus. These signals can be differentiated through additional preparatory steps. For instance, a similar narrative perspective and similar plot elements account for smaller distances between texts along with authorship, epoch, and sometimes even the protagonist's social class.

Further, we observed that pronouns were strongly associated with the narrator's perspective. Therefore, if this perspective is not relevant for the main research question, it may be necessary to undertake alterations to the text surrogate to eliminate this signal. Similarly, toponyms are candidates for alterations: if a project aims to

37 E.g. Matthew L. Jockers and Rosamond Thalken, *Text Analysis with R for Students of Literature*, 2nd ed. (Springer, 2020), https://doi.org/10.1007/978-3-030-39643-5.

examine the hierarchy between geographical places in rural settings, the toponym could be changed to a broader type of place.

It is important to keep in mind that signals discovered so far are specific to the *Deutscher Novellenschatz*-corpus. In subsequent research, detailed analyses of the other collections are planned. This order of exploration is motivated by the historical dimension of the collections: the *Deutscher Novellenschatz* is considered a genre-defining collection in literary historiography, and we expect our findings about the novella genre to be solidified on the basis of results related to the *Neuer Deutscher Novellenschatz*, thus further exploring the question whether the *Deutscher Novellenschatz*-collection can be considered a representative corpus for novellas of the 19th century.[38] As this article demonstrates, reading at scale links digital methods and operationalizations to the middle-sized *Novellenschatz*-corpus as a historical artefact, gradually including insights from data analysis and research in literary history in the adaptations of operationalizations. The concept allows switching between abstract representations of literary texts (such as word vectors and lists of most frequent words) and analytical text interpretations in the hermeneutic tradition, thus bringing both into a fruitful exchange.

Bibliography

Burrows, John. "'Delta': A Measure of Stylistic Difference and a Guide to Likely Authorship". *Literary and Linguistic Computing* 17, no. 3 (1 September 2002): 267–87. https://doi.org/10.1093/llc/17.3.267.

Büttner, Andreas, Friedrich Michael Dimpel, Stefan Evert, Fotis Jannidis, Steffen Pielström, Thomas Proisl, Isabella Reger, Christof Schöch, and Thorsten Vitt. "»Delta«in der stilometrischen Autorschaftsattribution". *Zeitschrift für digitale Geisteswissenschaften*, 2017. https://doi.org/10.17175/2017_006.

Erlin, Matt, and Lynne Tatlock. "Introduction: 'Distant Reading' and the Historiography of Nineteenth Century German Literature". In *Distant Readings: Topologies of German Culture in the Long Nineteenth Century*, edited by Matt Erlin and Lynne Tatlock, 1–29. Studies in German Literature, Linguistics, and Culture. Rochester, New York: Camden House, 2014.

Gervinus, Georg Gottfried. *Handbuch Der Geschichte Der Poetischen National-Literatur Der Deutschen*. 3. Aufl. Leipzig: Engelmann, 1844.

Ghannay, Sahar, Benoit Favre, Yannick Estève, and Nathalie Camelin. "Word Embedding Evaluation and Combination". In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, 300–305. Portorož, Slovenia:

38 Fotis Jannidis, "Perspektiven quantitativer Untersuchungen des Novellenschatzes", *Zeitschrift für Literaturwissenschaft und Linguistik* 47, no. 1 (March 2017): 7–27, https://doi.org/10.1007/s41244-017-0050-x.

European Language Resources Association (ELRA), 2016. https://aclanthology.org/L16-1046.

Gius, Evelyn, Jan Christoph Meister, Malte Meister, Marco Petris, Christian Bruck, Janina Jacke, Mareike Schumacher, Dominik Gerstorfer, Marie Flüh, and Jan Horstmann. "CATMA 6", 6 April 2022. https://doi.org/10.5281/ZENODO.1470118.

Havrlant, Lukáš, and Vladik Kreinovich. "A Simple Probabilistic Explanation of Term Frequency-Inverse Document Frequency (Tf-Idf) Heuristic (and Variations Motivated by This Explanation)". *International Journal of General Systems* 46, no. 1 (2017): 27–36. https://doi.org/10.1080/03081079.2017.1291635.

Heintz, Bettina. "Numerische Differenz. Überlegungen zu einer Soziologie des (quantitativen) Vergleichs / Numerical Difference. Toward a Sociology of (Quantitative) Comparisons". *Zeitschrift für Soziologie* 39, no. 3 (1 January 2010). https://doi.org/10.1515/zfsoz-2010-0301.

Hengchen, Simon, Ruben Ros, Jani Marjanen, and Mikko Tolonen. "A Data-Driven Approach to Studying Changing Vocabularies in Historical Newspaper Collections". *Digital Scholarship in the Humanities* 36, no. Supplement 2 (5 November 2021): 109–26. https://doi.org/10.1093/llc/fqab032.

Heyse, Paul, and Hermann Kurz. "Einleitung". In *Deutscher Novellenschatz*, Vol. 1. München: R. Oldenbourg, 1871. https://www.deutschestextarchiv.de/heysekurz_einleitung_1871.

Horstmann, Jan. "Undogmatic Literary Annotation with CATMA. Functions, Differentiation, Systematization". In *Annotations in Scholarly Editions and Research*, edited by Julia Nantke and Frederik Schlupkothen, 157–76. De Gruyter, 2020. https://doi.org/10.1515/9783110689112-008.

Jannidis, Fotis. "Perspektiven quantitativer Untersuchungen des Novellenschatzes". *Zeitschrift für Literaturwissenschaft und Linguistik* 47, no. 1 (March 2017): 7–27. https://doi.org/10.1007/s41244-017-0050-x.

Jockers, Matthew L. *Macroanalysis: Digital Methods and Literary History*. Topics in the Digital Humanities. Urbana, IL: University of Illinois Press, 2013.

Jockers, Matthew L., and Rosamond Thalken. *Text Analysis with R for Students of Literature*. 2nd ed. Springer, 2020. https://doi.org/10.1007/978-3-030-39643-5.

Kleymann, Rabea. "Datendiffraktion: Von Mixed zu Entangled Methods in den Digital Humanities". HTML,XML,PDF. Edited by Manuel Burghardt, Lisa Dieckmann, Timo Steyer, Peer Trilcke, Niels Walkowski, Joëlle Weis, and Ulrike Wuttke. *Fabrikation von Erkenntnis – Experimente in den Digital Humanities (Zeitschrift für digitale Geisteswissenschaften / Sonderband)*, 2022. https://doi.org/10.17175/SB005_008.

Krautter, Benjamin, Janis Pagel, Nils Reiter, and Marcus Willand. "Eponymous Heroes and Protagonists – Character Classification in German-Language Dramas". *LitLab Pamphlet #7* (2018): 58.

Lorenzini, Matteo, Marco Rospocher, and Sara Tonelli. "On Assessing Metadata Completeness in Digital Cultural Heritage Repositories". *Digital Scholarship in the Humanities* 36, no. Supplement 2 (5 November 2021): 182–88. https://doi.org/10.1093/llc/fqab036.

Manning, Christopher D., Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. "The Stanford CoreNLP Natural Language Processing Toolkit". In *Association for Computational Linguistics (ACL) System Demonstrations*, 55–60, 2014. http://www.aclweb.org/anthology/P/P14/P14-5010.

Moretti, Franco. "Conjectures on World Literature". *New Left Review* 1 (2000): 54–68.

Moretti, Franco. "Conjectures on World Literature". In *Distant Reading*, 43–62. London; New York: Verso, 2013.

Moretti, Franco. *Graphs, Maps, Trees: Abstract Models for a Literary History*. London; New York: Verso, 2005.

Mueller, Martin. "Morgensterns Spectacles or the Importance of Not-Reading". *Northwestern University Digital Humanities Laboratory* (blog), 2013. https://sites.northwestern.edu/nudhl/?p=433.

Mueller, Martin. *Scalable Reading*, 2019. https://scalablereading.northwestern.edu/.

Päpcke, Simon, Thomas Weitin, Katharina Herget, Anastasia Glawion, and Ulrik Brandes. "Stylometric Similarity in Literary Corpora: Non-Authorship Clustering and 'Deutscher Novellenschatz'". *Digital Scholarship in the Humanities*, 38, no. 1 (2023): 277–95, 1–19. https://doi.org/10.1093/llc/fqac039.

Piper, Andrew. *Can We Be Wrong? The Problem of Textual Evidence in a Time of Data*. Cambridge Elements. Digital Literary Studies. Cambridge: Cambridge University Press, 2020. https://doi.org/10.1017/9781108922036.

Piper, Andrew. *Enumerations: Data and Literary Study*. Chicago, IL; London: The University of Chicago Press, 2018.

Reiter, Nils, Anette Frank, and Oliver Hellwig. "An NLP-Based Cross-Document Approach to Narrative Structure Discovery". *Literary and Linguistic Computing* 29, no. 4 (January 2014): 583–605. https://doi.org/10.1093/llc/fqu055.

Rettig, Laura, Regula Hanggli, and Philippe Cudre-Mauroux. "The Best of Both Worlds: Context-Powered Word Embedding Combinations for Longitudinal Text Analysis". In *2020 IEEE International Conference on Big Data (Big Data)*, 4741–50. Atlanta, GA, USA: IEEE, 2020. https://doi.org/10.1109/BigData50022.2020.9377955.

Rybicki, J., D. Hoover, and M. Kestemont. "Collaborative Authorship: Conrad, Ford and Rolling Delta". *Literary and Linguistic Computing* 29, no. 3 (1 September 2014): 422–31. https://doi.org/10.1093/llc/fqu016.

Scherer, Wilhelm. *Geschichte Der Deutschen Litteratur*. 3. Auflage. Berlin: Weidmannsche Buchhandlung, 1885.

Scherer, Wilhelm, and Oskar Walzel. *Geschichte Der Deutschen Literatur.* Berlin: Askanischer Verlag, 1921.

Schöch, Christof. "Big? Smart? Clean? Messy? Data in the Humanities". *Journal of Digital Humanities* 2 (2013): 1–12.

Storm, Theodor. "Eine zurückgezogene Vorrede aus dem Jahre 1881". In *Theodor Storms sämtliche Werke in acht Bänden*, edited by Albert Köster, 8:122–23, 2012.

Thai, VinhTuan, and Siegfried Handschuh. "Context Stamp: A Topic-Based Content Abstraction for Visual Concordance Analysis". In *Proceedings of the 2011 Annual Conference Extended Abstracts on Human Factors in Computing Systems – CHI EA '11*, 2269. Vancouver, BC, Canada: ACM Press, 2011. https://doi.org/10.1145/1979742.1979906.

Walkhoff, Monika. *Der Briefwechsel Zwischen Paul Heyse Und Hermann Kurz in Den Jahren 1869 – 1873 Aus Anlass Der Herausgabe Des Deutschen Novellenschatzes [Mit Faks.]*. München: Foto-Druck Frank, 1967.

Walzel, Oskar. *Die Deutsche Dichtung Seit Goethes Tod*. Berlin: Askanischer Verlag, 1919.

Weitin, Thomas. "Average and Distinction. The Deutsche Novellenschatz Between Literary History and Corpus Analysis", LitLab Pamphlet, 6 (2018): 1–23.

Weitin, Thomas. "Burrows's Delta Und Z-Score-Differenz Im Netzwerkvergleich. Analysen Zum Deutschen. Novellenschatz von Paul Heyse Und Hermann Kurz (1871–1876)". In *Digitale Literaturwissenschaft. Beiträge Des DFG-Symposiums 2017*, edited by Fotis Jannidis. Germanistische Symposien. Stuttgart: Metzler, 2023.

Weitin, Thomas. *Digitale Literaturgeschichte. Eine Versuchsreihe in 7 Experimenten*. Berlin: Metzler / Springer Nature, 2021.

Weitin, Thomas. "Thinking Slowly. Reading Literature in the Aftermath of Big Data". *LitLab Pamphlet* #1 (2015): 1–19.

Weitin, Thomas, and Katharina Herget. "Falkentopics: Über Einige Probleme Beim Topic Modeling Literarischer Texte". *Zeitschrift Für Literaturwissenschaft Und Linguistik* 47, no. 1 (2017): 29–48. https://doi.org/10.1007/s41244-017-0049-3.

Corpora

Weitin, Thomas. Digitalized text corpus. 'Der Deutsche Novellenschatz'. Edited by Paul Heyse, Hermann Kurz. 24 volumes, 1871–1876. Darmstadt/Konstanz, 2016. https://www.deutschestextarchiv.de/doku/textquellen#novellenschatz.

Weitin, Thomas and Herget, Katharina. Digitalized text corpus. 'Der Neue Deutsche Novellenschatz'. Edited by Paul Heyse, Ludwig Laistner. 24 volumes, 1884–1887. Darmstadt, 2022. DOI: 10.5281/zenodo.6783577.

Weitin, Thomas and Herget, Katharina. Digitalized text corpus. 'Novellenschatz des Auslandes'. Edited by Paul Heyse, Hermann Kurz. 14 volumes, 1877–1884. Darmstadt, 2022. DOI: 10.5281/zenodo.6784080.

#CORPUS

Concerning the term #Corpus, the Oxford English Dictionary already sketches a conceptual extension towards analysis, a means for comparison. The well-established humanistic meaning of a "body or complete collection of writings or the like" or "the whole body of literature on any subject" dating back to the 18th c, was extended in mid-20th c by the understanding that it might represent a "material upon which a linguistic analysis is based" and by the emergence of corpus linguistics as a "branch of linguistics concerned with analysis of corpora as a means of studying language". *

From this, it is only a small step to the application of digital approaches to fields such as digital linguistics or digital literary studies although the challenge of defining a perfect or even but suitable #Corpus remains. The projects of mixed methods take it for granted and discuss the concept of corpus in two further directions. On the one hand, they extend it unselfconsciously from text and language towards music, visual media and even material objects. On the other hand, the terminus #Corpus is at the centre of diverse methodical considerations. Some projects focus on source material from humanities itself and its usability for digital analysis (Handwriting; ArchiMediaL; Rhythmicalizer; Digital Plato), some on the intrinsic characteristics of a specific corpus (QuaDrama) or the specific transformation from analogous to digital corpus with the resulting promises of knowledge (ReadingAtScale).

Most projects address challenges at the intersection between humanities and computer science as a crucial phenomenon: "Material in corpora is often manually or automatically enriched (..) for conducting large-scale experiments and following specific research questions" (QuaDrama), needs translation into machine-readable formats (BachBeatles) or additional means of sourcing to create an adequate data base (ArchiMediaL; see #Heterogeneity)

* "Corpus, n.". in: *Oxford English Dictionary (OED)*, first published 1893; most recently modified version published online March 2022 with draft additions from 1997 and 2013, https://www.oed.com/ [accessed: 20.05.2022].

Title: Doing DH in a mixed-methods scenario. Experiences in the QuaDramaA Project
Team: Quantitative Drama Analytics (Nils Reiter, Marcus Willand, Benjamin Krautter, Janis Pagel
Corpus: German Drama (1730–1930)
Field of Study: Literary Studies
Institution: University of Stuttgart
Methods: Annotation, machine learning, quantitative analysis
Tools: CorefAnnotator, DramaAnalysis, DramaNLP, RStudio
Technology: Apache UIMA, NLP, XML

On Designing Collaboration in a Mixed-Methods Scenario. Reflecting Quantitative Drama Analytics (QuaDrama)

Janis Pagel, Benjamin Krautter, Melanie Andresen, Marcus Willand, Nils Reiter

Abstract *'Quantitative Drama Analytics' (QuaDramA) is a mixed-methods project that brings together researchers from modern German literature and computational linguistics, thus belonging to the field of computational literary studies. The goal of the project is to define, annotate, automatically detect, and quantitatively analyze different dramatic character types in German-language plays. To this end, we extract textual and structural properties from plays and investigate their distribution among literary characters, such as Romeo and Juliet. One of the key decisions in any mixed-methods project is to agree on a specific collaboration workflow between the different parties, yet this decision is rarely made deliberately and explicitly. In QuaDramA, the collaboration is based on three pillars. (i) Text annotation is used to both clarify concepts and enrich data required in machine learning. (ii) The different project parts perform close and frequent personal collaboration, particularly in the beginning of the project. In addition, quantitative analyses are made by researchers from both disciplines by using generic programming tools (thus avoiding the need for custom graphical user interfaces). (iii) Quantitative and qualitative research methods go hand in hand and are closely integrated. We discuss the reasoning behind these pillars in this article and provide some recommendations for projects with a similar setup.*

Introduction

This article focuses on organization and collaboration in 'Quantitative Drama Analytics' (QuaDramA), a project between researchers from literary studies and computational linguistics (CL).[1] QuaDramA aims to extend the possibilities for large- and small-scale analysis of plays by using quantitative and formal methods, while focusing on dramatic characters as objects of investigation.

1 The scientific contributions of the project can be found on its web page: https://quadrama.github.io.

One of the biggest challenges in this collaboration is that research questions and presumptions from literary studies cannot be put into a computer directly, as they are usually context-dependent, example-based, and not aimed at quantification. In other words, they are defined in a less formalised way than would be needed for computational processing.[2] This gap between notions in literary studies and technical possibilities needs to be bridged by means of operationalization: Concepts, questions and phenomena need to be made measurable.[3] This step involves conceptual decisions and, therefore, it is more than a simple translation process. In addition, the results that are generated using algorithmic methods need to be 're-mapped' into the field of literary studies. The challenge of operationalization has two directions: From the humanities' conceptual realm into the formal realm and back. We postulate that both directions are interpretative processes. An algorithm formalizing the detection of a type of text is an interpretation of this type of text; mapping quantitative results (that are generated, e.g., in the form of bar charts) to humanities findings requires not only an interpretation of the numbers and their visual representation, but also simultaneous consideration of relevant contexts. These processes can fail for numerous reasons. We argue, however, that many of these reasons do not lie in technical or conceptual issues, but are rather organizational, pragmatic, or communication problems.

Operationalization, in this sense, can be thought of as being hierarchical: A target concept is operationalized in terms of more basic concepts that, in turn, are operationalized in even more basic concepts. Take, for instance, the 'protagonist' or 'main character' of a play. Operationalization of this concept consists of a definition of properties that are relevant for detecting a protagonist, such as the importance of the character for the plot, its relations to other characters or the themes they talk about. These properties themselves need to be operationalized: Themes in character speech could be operationalized through topic modeling or word field analysis, character relations could be operationalized through social networks. This way, an operationalization hierarchy is established, thus connecting the target concept to the textual surface.

Due to this hierarchical structure, operationalizations also provide information on a more fine-grained level: investigating the distribution of properties 'below' the

2 Cf. Jan Christoph Meister, "Computerphilologie vs. 'Digital Text Studies'", in: Christine Grond-Rigler and Wolfgang Straub (eds.): *Literatur und Digitalisierung* (Berlin, Boston: De Gruyter, 2013), 294.

3 See Axel Pichler and Nils Reiter, "Reflektierte Textanalyse", in: Nils Reiter, Axel Pichler, and Jonas Kuhn (eds.), *Reflektierte Algorithmische Textanalyse* (Berlin, Boston: De Gruyter, 2020), 45–47 and Axel Pichler, and Nils Reiter, "Zur Operationalisierung literaturwissenschaftlicher Begriffe in der algorithmischen Textanalyse: Eine Annäherung über Norbert Altenhofers hermeneutischer Modellinterpretation von Kleists *Das Erdbeben in Chili*", *Journal of Literary Theory* 15, no. 1–2 (2021): 4–7.

actual target concept may be just as revealing as looking at the target concept itself. Instead of only knowing that a character is a protagonist, we are *also* aware of the (operationalized) underlying property values and can compare them:[4] the number of words a character utters, the number of relations they have in a copresence network,[5] or the distribution of different topics they speak about. This feature-based description is more specific and tries to be more transparent than the holistic and interpretative notion of 'protagonist' as it is used in literary studies; as a result, it provides new opportunities to pose research questions.[6]

Summarizing our experiences in QuaDramA, we argue that digital humanities (DH) projects work better if the interdisciplinary collaboration is thought through in detail and constantly evaluated as well as adapted. We, therefore, describe the three pillars that our collaboration relies on and give insight into our reasoning behind them. In the next section, we provide an overview of the project itself and its research agenda. We then discuss the three pillars individually: (i) annotation as a means to enable interdisciplinary discussion, (ii) our collaboration model, i.e., the setup of our tools and workflows, and (iii) a close integration of quantitative and qualitative research. Finally, we provide conclusions.

Quantitative Drama Analytics (QuaDramA)

The project QuaDramA develops methods for the quantitative analysis of dramatic texts. The methods are integrated into a research framework that combines the computing power of digital tools in CL with established qualitative methods of literary

4 For instance, in Lessing's *Minna von Barnhelm* (1767) Minna's maid Franziska is the play's character who is most frequently present on stage: she appears and speaks in 32 of the 56 scenes. The title character Minna, however, speaks only in 25 scenes. Still, Franziska would not be regarded as the protagonist of the play. A quantitative indication for this is Franziska speaking a lot less than Minna. Franziska's character speech comprises only 3,904 tokens, while Minna's comprises 7,347. See also Manfred Pfister, *The Theory and Analysis of Drama*, transl. John Halliday (Cambridge et al.: Cambridge University Press, 1988), 165–66.

5 Copresence networks are networks based on literary characters that are on stage at the same time. Such social networks have many different applications (see Fazli Can, Tansel Özyer and Faruk Polat (eds.), *State of the Art Applications of Social Network Analysis* (Cham: Springer, 2014)), but their use is not without pitfalls. See Katharina A. Zweig, *Network Analysis Literacy. A Practical Approach to the Analysis of Networks* (Wien: Springer, 2016). One might expect protagonists to have a central position in the network, interacting with many of the other characters.

6 It may not be directly obvious that exactness is achievable or even desirable when applied to humanities concepts. We argue that this depends on the specific context in which such concepts are used. Concepts that are primarily descriptive need to be as exact as possible, while the same does not necessarily hold true for interpretations of literary texts.

studies. In this way, we gain a better understanding of drama history, its contexts, and the interpretation of individual plays. While structure-oriented drama research is straightforward, and has been conducted extensively, e.g., in the form of network analysis going back to the early 1970s,[7] QuaDramA combines the analysis of the structure of the plays with that of their content.

The corpus we are investigating is the German Drama Corpus (GerDraCor)[8] which comprises more than 590 German plays, mainly from 1730 to 1930. Our core method to generate insight is a statistical view on the literary data, properly interpreted with the background knowledge that has been established in literary studies, and, if possible, used as a backdrop for traditional close reading. The project, thus, is 'mixed-methods' in the sense that quantitative and hermeneutic methods are intertwined and used side by side, consequently allowing us to evaluate their results in comparison. As not all relevant textual or structural properties are directly accessible on the text surface, we employ methods from computational linguistics—based on machine learning or hand-crafted rule sets—to assign relevant properties such as themes characters talk about on a large scale. In addition to this use as a pre-processing step, machine-learning techniques are used to gain insight into the target categories, e.g., the different character types in a play. This is done by inspecting important features in a classification system or by doing an in-depth error analysis of automatic systems.

In comparison to traditional literary studies, quantitative methods add a new layer of analysis to the spectrum of hermeneutical reading. Although a pluralism of methods does exist in literary studies, these methods are to a great extent qualitative and derive from theories focusing on different entities (author, text, or reader) as their main reference for interpretation. For the practices of literary studies, quantitative methods and the emerging data-based text knowledge can support the understanding of single literary texts and, on a larger scale, literary history, including the formation of literary movements, genres, or conditions of production and reception.

Computational linguistics, the D-discipline in this DH-collaboration, is in general concerned with detecting linguistic properties in texts (and other forms of human language expressions). CL has made tremendous progress in the past 10 years, mostly due to the availability of large data sets, efficient hardware, and machine-learning algorithms that make use of improved computing power (in particular, artificial neural networks). At the same time, most of the progress in CL is made on

7 See Solomon Marcus, *Mathematische Poetik*, transl. Edith Mândroiu (Frankfurt a.M.: Athenäum Verlag, 1973), 287–370.

8 Frank Fischer et al., "Programmable Corpora. Die digitale Literaturwissenschaft zwischen Forschung und Infrastruktur am Beispiel von DraCor", in *Abstracts of DHd*, ed. Patrick Sahle (Frankfurt, Mainz, 2019), 194–197, https://doi.org/10.5281/zenodo.2596095.

English newspaper texts, because 'news' is considered to be the default domain and English a world language.[9] While applying CL tools on other textual domains, the drop in their performance can be quite significant,[10] and other languages are much less supported. While the progress is impressive and the result of hard work, developing CL methods only on newspaper texts leads to models that lack linguistic properties because they simply do not exist in this genre. For the processing of (German) dramatic texts, this applies to the way plays are layered into uttered tokens, stage directions and other kinds of 'paratext' (segmentation markers, speaker designations, etc.). Crucially, these three layers cannot be considered as three separate texts and processed independently, because a number of linguistic relations can be discerned between them. The excerpt below, taken from Lessing's Emilia Galotti (1772), gives an example for this phenomenon. The underlined words are part of a coreference chain[11] that starts within the Prince's utterance, but continues within the stage direction associated with Conti, and is then referenced by Conti in his utterance. It shows a linguistic continuity that cuts across two different textual layers, also revealing that a proper linguistic modeling of these texts needs to take the different layers and the specificities of the text genre into account.

As we hope has become clear, the research questions in QuaDramA require a close integration of literary studies and computational linguistics, and therefore a close collaboration between researchers of these disciplines. The next sections describe guiding principles that have made this collaboration constructive for us.

[Prince and Conti discuss two paintings that Conti painted]
***Prince**. 'Tis true, but why did you not bring it a month sooner? Lay it aside. What is the other?*
***Conti**. (taking it up and holding it still reversed). It is also a female portrait.*[12]

9 The focus on a single domain has enabled the recent boost in prediction quality because it keeps one 'variable' controlled while doing experiments. Another reason why news (and, increasingly, social media texts) are in focus is that the amount of text in need of processing increases daily and will continue to do so. This justifies a large investment of resources, because every increase in performance will pay off in the long run. The situation is very different if the text corpus is limited and cannot be expected to grow in the future.

10 See Nils Reiter, *Discovering Structural Similarities in Narrative Texts using Event Alignment Algorithms* (PhD diss., Heidelberg University, 2014), https://doi.org/10.11588/heidok.00017042 and Benedikt Adelmann et al., "Evaluating Part-of-Speech and Morphological Tagging for Humanities' Interpretation", in: Andrew U. Frank et al. (eds.): *Proceedings of the Second Workshop on Corpus-Based Research in the Humanities* (Wien, 2018), 5–14, https://www.oeaw.ac.at/fileadmin/subsites/academiaecorpora/PDF/CRH2.pdf.

11 In coreference annotation, every mention of an entity (e.g., person, object, abstract concept) is marked in the text, be it a proper name, a noun phrase or a pronoun.

12 Taken from Lessing's Emilia Galotti (1772). The excerpt shows a coreference chain (underlined) that crosses the text layers (transl. by Ernest Bell, 1878).

Collaboration Model

1. Text Annotation

In a mixed-methods scenario, many different traditions, methods, and backgrounds meet and need to interact in a meaningful and efficient way. For QuaDramA, it means finding a way to collaborate between researchers from modern German literary studies and CL. One pillar of our solution is to use manual *annotation* as a layer of communication[13] that allows us to talk about data, concepts, theories, and, of course, elements of literary texts such as character types.

Annotation is the process of enriching data with overt information in the form of markup, which makes underlying, implicit information[14] explicit.[15] In CL, annotation and the use of annotated data has a long tradition.[16] E.g., computational linguists will mark all occurrences of words in a text and assign a specific part of speech to each word. These markups have (at least) two benefits: (i) The created data can be used to inform and train automatic models, and (ii) the process of annotation discloses difficult cases for assigning markup (e.g., deciding which part of speech a certain word should receive) and thus sparks discussions about the underlying theory and guidelines used to annotate the phenomenon in question. It relates to the notion of operationalization that we have outlined above. In order to create sensible annotations, the target concepts need to be operationalized. At the same time, annotation is one possible option to make the target concepts measurable. This problem of interdependence can be solved by iteratively annotating data and including the insights from the annotations into new or refined operationalizations which in turn inform new annotation decisions.

---

13 An observation that we have made is that in order to establish a healthy communication between different disciplines, a chicken-and-egg problem needs to be solved: On the one hand, annotations are a gateway for developing a common language but, on the other, they are a realm for which a common language is a prerequisite. This problem can be solved by first creating annotations 'blindly' and then using them to establish mutual viewpoints. In this sense, annotations are a concrete object for the two parties to refer to while discussing phenomena and assumptions about the texts.

14 Implicit information can be located relatively overtly on the text surface, e.g., when appearances and exits of characters are annotated in a play. Usually, however, implicit information is not directly part of the text, but requires theoretical assumptions which provide categories that can then be mapped to textual properties, e.g., when the sentiments of character speech are annotated.

15 See James Pustejovsky and Amber Stubbs, *Natural Language Annotation for Machine Learning: A Guide to Corpus-Building for Applications* (Sebastopol/Boston/Farnham: O'Reilly Media, 2012).

16 See Eduard Hovy and Julia Lavid, "Towards a Science of Corpus Annotation: A New Methodological Challenge for Corpus Linguistics", *International Journal of Translation Studies* 22, no. 1 (2010): 13–36.

In our project, we realize this iterative process in the following, partially overlapping steps: Firstly, we use the process of annotation to condense the knowledge of literary studies about certain aspects of dramatic theory, and thus make it more concrete. This approach of concretization is used in numerous other DH projects.[17] It can also be exemplified with our efforts to annotate protagonists of plays.[18] We collected theories and characteristics of what it means to be a play's protagonist drawing on both primary sources (e.g., Aristotle's *Poetics*) and secondary literature. Next, we condensed the given descriptive observations or normative instructions into an annotation framework, such that independent annotators could mark characters of plays by using our guidelines. In doing so, we were able to establish which properties protagonists should share most of the time. Based on the annotated data we can start to investigate the different properties that can be individually identified in a given text. How much do protagonists talk? Who do they talk to? What do they talk about? Deriving from general and abstract conceptualizations, the annotation allows us to directly compare characters that we have identified as protagonists. Finally, the resulting data, i.e., the information about which character in a play can be classified as a protagonist by which features, can be used to conduct further research, e.g., by automatically classifying protagonists and determining which properties of a character are most useful for classification decisions. Furthermore, it is possible to use the annotations as input for other classification tasks, such as identifying character types.[19]

Text annotations can be represented digitally in a multitude of ways. We opted for two main formats that allow us to cover the needs of the two research areas in QuaDramA: TEI and CoNLL. The TEI[20] format is an XML[21] standard that describes a rich catalog of categories for encoding and marking surface and semantic aspects of texts. It is popular in the DH and computational literary studies community and as such perfectly suited our purpose. A well-known format in CL that is used to encode

17 See Evelyn Gius and Janina Jacke. "The Hermeneutic Profit of Annotation: On Preventing and Fostering Disagreement in Literary Analysis", *International Journal of Humanities and Arts Computing* 11, no. 2 (2017): 233–254, https://doi.org/10.3366/ijhac.2017.0194 and Nils Reiter, Marcus Willand and Evelyn Gius, "A Shared Task for the Digital Humanities Chapter 1: Introduction to Annotation, Narrative Levels and Shared Tasks", *Cultural Analytics* (2019), https://culturalanalytics.org/article/11192

18 See Benjamin Krautter et al., "Titelhelden und Protagonisten – interpretierbare Figurenklassifikation in deutschsprachigen Dramen", *LitLab Pamphlets* 7 (2018), https://www.digitalhumanitiescooperation.de/wp-content/uploads/2018/12/p07_krautter_et_al.pdf.

19 See Krautter et al., "Titelhelden und Protagonisten".; Janis Pagel et al., "Annotation als flexibel einsetzbare Methode", in: Nils Reiter, Axel Pichler, and Jonas Kuhn (eds.): *Reflektierte Algorithmische Textanalyse* (Berlin, Boston: De Gruyter, 2020), 125–141.

20 Text Encoding Initiative, https://tei-c.org/.

21 Extensible Markup Language.

linguistic annotations is the CoNLL[22] format. It has been used for many shared tasks and constitutes a column-based format where each row represents a token and the columns represent linguistic information such as parts of speech, lemma, or syntactic information. As we can convert information between these formats and other formats if needed, we ensure that our annotations can be used by a variety of different communities and for different purposes.

One type of annotation that poses unique challenges is that of coreference (see Figure 1 for an example). The annotation of coreference is often performed on short texts such as newspaper or Wikipedia articles. Dramatic texts, however, are much longer and contain far more entities, which makes existing workflows for annotating coreference unapplicable. Thus, we created our own annotation tool, CorefAnnotator,[23] that allowed us to manually perform coreference annotation on longer texts and to display the texts in a suitable format. We also established our own guidelines and a specific workflow.[24] The annotation tool CorefAnnotator supports the import of dramatic texts directly from TEI/XML and can export both to the aforementioned CoNLL format and TEI/XML. Internally, CorefAnnotator uses the Apache UIMA framework to represent annotations.[25]

Coreference annotation also reveals ambiguities and uncertainties with respect to literary interpretation. Oftentimes, several annotations are valid since the text itself offers multiple possible readings. These cases are to be distinguished from 'false' ambiguities that are caused by annotation mistakes or improper guidelines.[26] In the case of valid ambiguities, we ask annotators to mark the different possibilities.

We have published a first version of a corpus with coreference annotations on dramatic texts,[27] called *GerDraCor-Coref*. The corpus release allows us to document a milestone within the project and enables further subsequent research, as the corpus can now be cited and used as a resource for various investigations, such as personality traits of characters or the function of abstract entities in drama. As mentioned

22 Named after the 'Conference on Computational Natural Language Learning', which first used it.

23 https://github.com/nilsreiter/CorefAnnotator.

24 See Ina Rösiger, Sarah Schulz and Nils Reiter, "Towards Coreference for Literary Text: Analyzing Domain-Specific Phenomena", in: *Proceedings of the Second Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature* (Santa Fe, 2018), 129–38, http://aclweb.org/anthology/W18-4515.

25 For a detailed description, see Nils Reiter, "CorefAnnotator – a New Annotation Tool for Entity References", in: *Abstracts of EADH: Data in the Digital Humanities* (Galway, 2018), https://doi.org/10.18419/opus-10144.

26 See Gius and Jacke, "The Hermeneutic Profit of Annotation".

27 Janis Pagel, and Nils Reiter, "GerDraCor-Coref: A Coreference Corpus for Dramatic Texts in German", in: *Proceedings of the 12th Language Resources and Evaluation Conference (LREC)* (Marseille, 2020), 55–64, https://www.aclweb.org/anthology/2020.lrec-1.7.pdf.

before, plays contain much longer coreference chains on average, which we can now quantify in our corpus (328.8 mentions in a coreference chain on average, compared to 14.4 in newspaper data[28]). This shows the value that annotations as a stand-alone task have while simultaneously fostering further research.

2. Workflow

Organizing an interdisciplinary collaboration on a daily basis is never easy, as all DH-experienced researchers are aware of. The collaboration model in QuaDramA consists of a technical and a non-technical part, and both have been carefully and consciously developed. This section discusses the core ideas of our collaboration model, namely to a) focus on programming libraries instead of tools, b) modularize the technical workflow such that its steps have clearly defined interfaces, and c) provide external tutorials and workshops.

In early stages, we envisaged an interactive analysis tool with a graphical user interface and made some experiments in that direction. It quickly turned out to be unfeasible: The significant development time between 'having an idea for an analysis' and 'inspecting the results of this analysis' made an explorative approach quite difficult and, at the same time, prohibited serious research on the technical side because the computational linguists in the project would be busy developing and refining the user interface all the time. For the scholars in the humanities, this would also lead to fundamental dependence: They would never be able to conduct analyses on their own and therefore would always need help from a 'technician'. This experience led us to drop the tool idea entirely. Instead, we focused on programming libraries.

A programming library can be described as a collection of functions that work well together for a clearly defined and limited functional area.[29] To use them, one needs to write programming code to execute the functions and do something with the results. Most of the analysis functions we are using on dramatic texts are collected in the library 'DramaAnalysis'[30], developed for the programming language called 'R'. The library is available on the standard repository for R packages (Comprehensive R Archive Network, CRAN) which makes the installation quite easy. In

28 Since newspaper texts are usually much shorter than plays, we have also normalized the values by text length, but are still getting higher values for plays on average (0.0089 for plays, 0.0013 for newspaper data).

29 For instance, the Python library 'scikit-learn' provides functions for machine learning, the Java library 'Apache Commons CSV' provides functions for handling CSV files etc. Programming libraries are an essential part of programming and complex programs may use hundreds or thousands of libraries.

30 http://doi.org/10.5281/zenodo.597509.

addition to our own project, the package is also used by others to analyze and visualize dramatic texts.[31]

Functions provided by DramaAnalysis can easily be used in RStudio, an integrated development environment (IDE) and a data science platform for the R programming language. Basic knowledge of the programming language R is, however, required to use it: The main code that calls the functions of the package and sets its parameters still needs to be written by the user. We believe that it is an acceptable requirement since it is a skill that will benefit the non-programmers beyond the project, making them more independent in the long run.[32] To soften the start, the package provides documentation with examples, a report function to easily generate a comprehensive report for a single play, and a 100-page tutorial with examples, guides and interpretation hints.[33]

With respect to visualization, this setup allows one to make use of the broad visualization capabilities of R. As R is oriented towards data analysis and statistics, the visualization methods we use are bar charts, box and scatter plots, and node-link diagrams for networks. As they are generated on the fly using R code, they can be manipulated easily, offering a simple way of interactive visualization by adapting the code directly.

This shift from tools to libraries reflects an adjustment in our 'mental' setup: In contrast to a common, seemingly natural division of labor, we do not see literary scholars primarily as users and computational linguists as developers. In our experience, it amounts to collaboration which is equally beneficial for both parties.

Modularization is another concept that promotes independence among researchers. It refers to the idea that a software project is subdivided into modules, each with their own goal and ability to operate largely independent of another. The interaction between the modules is regulated by a clear definition of what the input and output of each module are, so that one module can further process the output of another. The technical setup in QuaDramA consists of two main modules, reflecting the research interests and goals of the involved parties. The first module includes the language processing and is mostly the working field of the computational linguists. The second is the subsequent quantitative analysis of dramatic texts and is mostly done by the literary scholars. Both modules and their interaction will be briefly described in the following.

---

31 See Hanno Ehrlicher et al., "Gattungspoetik in quantitativer Sicht. Das Werk Calderón de la Barcas und die zeitgenössischen Dramenpoetiken", *LitLab Pamphlet* 9 (2020): 1–29. https://www.digitalhumanitiescooperation.de/wp-content/uploads/2020/04/p09_ehrlicher_lehmann_al_de.pdf.

32 In contrast, a project-specific tool will quickly be disbanded once the project is finished and to continue to be familiar with it will be of little use.

33 https://quadrama.github.io/DramaAnalysis/tutorial/3/.

Our corpus is encoded in TEI-XML and provides a machine-readable structuring (in acts/scenes) as well as speaker designations and separable stage directions. In our pipeline, the texts are processed by the custom-built DramaNLP library[34] which is built on top of Apache UIMA.[35] All the information in the original XML files is initially mapped to UIMA data structures. The processing pipeline automatically adds linguistic information such as parts of speech, lemmas, syntactic parse trees, and coreference chains. As the latter is further developed within the project, its pipeline component is updated from time to time. The final component of DramaNLP is an export of the data into several comma-separated values (CSV) files which are stored in a Git repository on GitHub, containing information about the segmentation of the plays, the spoken tokens, the tokens in stage directions, the characters, and metadata about the play. The use of Git provides a straightforward versioning mechanism, such that earlier data versions and experiments based on them can be reproduced easily. The CSV files serve as an interface to the second module comprising the quantitative analysis with R.

The second part of the technical setup consists of the R library DramaAnalysis which reads the CSV files and offers analysis and visualization functions. In order to make access and navigation of the package's capabilities easier, we have created a tutorial which interested scholars can work through. Starting from loading a specified play (with the function loadDrama()), one can investigate themes in character speech by using word fields (with the function dictionaryStatistics()), access character statistics (function: characterStatistics()), and access utterance statistics (utteranceStatistics()). Also, one can inspect configurations (configuration(), that is, co-occurrence matrices),[36] get information on active and passive presence of characters (presence()),[37] calculate different metrics that measure the rate of stage personnel exchange over scene boundaries (hamming(), scenicDifference())[38] and calculate a general frequency matrix of words that can, for instance, be used in the stylometry library 'stylo'.[39]

---

34 http://doi.org/10.5281/zenodo.597866; https://github.com/quadrama/DramaNLP.

35 UIMA is a processing library that provides data structures and efficient indexing for handling annotations on textual data. https://uima.apache.org.

36 See Pfister, The Theory and Analysis of Drama, 171–176.

37 See Marcus Willand et al., "Passive Präsenz tragischer Hauptfiguren im Drama", in: Christof Schöch (ed.): *Abstracts of DHd* (Paderborn, 2020), 177–181, https://zenodo.org/record/3666690.

38 See Frank Fischer et al., "Network Dynamics, Plot Analysis. Approaching the Progressive Structuration of Literary Texts", in *Digital Humanities 2017. Conference Abstracts* (Montréal, 2017), 437–441, https://dh2017.adho.org/abstracts/DH2017-abstracts.pdf.

39 Maciej Eder, Jan Rybicki and Mike Kestemont, "Stylometry with R: A Package for Computational Text Analysis", R Journal 8, no. 1 (2016): 107–21. https://doi.org/10.32614/RJ-2016-007.

This modular structure of the project ensures that project partners can, for the most part, work independently of each other as long as the interface remains stable, and thus makes smooth collaboration possible.

In addition, we consider dissemination activities geared towards the 'outside' to be a part of the collaboration model that has worked well in our project. These activities are mostly workshops and tutorials held on various occasions.[40] While such activities are obviously beneficial for an external view on a project, they also serve an important internal goal: While preparing such an activity, the team is forced to condense the important aspects into a short period of time and to restructure it in a way that is useful for others. Thus, the preparation fosters the uncovering of misunderstandings and misconceptions, which is very useful for the project itself. Finally, such milestones also encourage everyone to provide proper documentation and code structures.

3. Integration of Quantitative and Qualitative Analysis

The following section discloses our reasoning in combining qualitative and quantitative methods for the analysis of drama. This is the third pillar of our collaboration and it ensures that our two disciplines work together, instead of next to each other, to benefit from each other's insights. The employment of quantitative methods has been hotly debated in literary studies and we, therefore, start by providing a brief overview of the ongoing methodological debate on distant and scalable reading. Thereafter, we describe three examples of our research that illustrate the integration of quantitative and qualitative methods in QuaDramA.

In the early 2000s, 'distant reading', a term first coined by Franco Moretti in his essay *Conjectures on World Literature* (2000), sparked a methodological controversy about the analysis of literary history. Moretti, in the light of his diagnosis that literary history should not solely rely on its canonical fracture,[41] but also on what Mar-

40 Tutorial at the Heidelberg summer school on quantitative drama analytics, 2019; Tutorial in a research seminar at Tübingen University, 2020; Operationalization workshop at the German DH conference (DHd), 2020 and 2022; Tutorial in a research seminar at the University of Bochum (2021); Operationalization workshop at the international DH conference, 2022.

41 The literary canon is a compilation of texts that are recognized as exemplary and are, therefore, considered particularly worthy of remembrance. Consequently, the canon is highly selective and excludes major parts of published literature. Moretti, for instance, projects that the literary canon of nineteenth-century British novels would consist 'of less than one per cent of the novels that were actually published'. Franco Moretti, "Graphs, Maps, Trees: Abstract Models for Literary History", *New Left Review* 24 (2003): 67. For researchers such as Matthew Jockers, relying on the literary canon to study literary history is not the way forward: 'Just as we would not expect an economist to generate sound theories about the economy by studying a few consumers or a few businesses, literary scholars cannot be content to read literary history from a canon of a few authors or even several hundred texts'. Matthew L.

garet Cohen called the "great unread",[42] polemically suggested embarking on "second hand" criticism instead of close reading of literature:[43]

> "the trouble with close reading (in all of its incarnations, from the new criticism to deconstruction) is that it necessarily depends on an extremely small canon. This may have become an unconscious and invisible premise by now, but it is an iron one nonetheless: you invest so much in individual texts *only* if you think that very few of them really matter. Otherwise, it doesn't make sense. And if you want to look beyond the canon [...] close reading will not do it."[44]

His proposition to do 'distant reading' back then was based on mostly two ideas that were rooted in his philosophical conviction of a materialistic theory of history: (i) reading secondary literature instead of literature itself, i.e., analyzing other researchers' analyses, and (ii) activating research networks with experts for different languages, genres and time spans.[45] His ambitious endeavor to analyze world literature, in the end, depends on merging as much existing knowledge as possible for as many literary texts as possible. Thus, the focus shifts from understanding and interpreting a single literary text to the explanatory power of patterns to explain literary conventions and their historical development.[46] While doing 'distant reading', "the reality of the text undergoes a process of deliberate reduction and abstraction"[47] for the bigger picture to emerge: 'it allows you to focus on units that are much smaller or much larger than the text: devices, themes, tropes—or genres and systems.'[48]

Back in 2000, Moretti neither referred to computational approaches, nor did he specifically mention quantitative methods to analyze literature. Later, in an essay called *Graphs, Maps, Trees* (2003), which became part of an essay collection with the same title, Moretti decidedly called for the construction of abstract models imported from theories of other disciplines, namely quantitative history, geography and evolutionary theory.[49] The resulting graphs, maps and trees have drawn a first connec-

---

Jockers, *Macroanalysis. Digital Methods and Literary History* (Urbana/Chicago/Springfield: University of Illinois Press, 2013), 9.

42 Margaret Cohen, *The Sentimental Education of the Novel* (Princeton: Princeton University Press, 1999), 23 and Margaret Cohen, "Narratology in the Archive of Literature", *Representations* 108, no. 1 (2009): 59.

43 Franco Moretti, "Conjectures on World Literature", *New Left Review* 1 (2000): 57.

44 Ibid.

45 See Ibid., 58–60.

46 See Franco Moretti, Graphs, Maps, Trees: Abstract Models for Literary History (London/New York: Verso, 2005), 91.

47 Ibid., 1.

48 Moretti, "Conjectures on World Literature", 57.

49 See Moretti, "Graphs, Maps, Trees", 2003, 67.

tion of 'distant reading' to computational methods that has deepened remarkably in the following years, mainly since 2010 when the Stanford Literary Lab was formed.[50]

Nowadays, 'distant reading' is pervasively used as a metaphor for computational quantitative corpus analyses.[51] However, its polemical introduction more than 20 years ago and its preference for reductionist models have since induced many critics to fundamentally object against analyzing art by calculating,[52] against the lack of innovative and new insights compared to the massive efforts needed,[53] and against the nature of computational methods from a perspective of intellectual work ethics: quantitative analysis of art was perceived as solely exploratory, as 'mere play'.[54] All these criticisms can be understood as central problems of the integration of quantitative and qualitative methods.[55]

Martin Mueller, a proponent of computational methods praising the "second order query potential of digital surrogates",[56] is also one of the critics of Moretti's 'distant reading' idea. For Mueller, 'distant reading' did not "express adequately the powers that new technologies bring to the old business of reading", as the term

> "implicitly set[s] the 'digital' into an unwelcome opposition to some other—a trend explicitly supported by the term 'Digital Humanities' or its short form DH,

50 See Ted Underwood, "The Stanford Literary Lab's Story", last modified February 11, 2017, https://www.publicbooks.org/the-stanford-literary-labs-narrative/https://www.publicbooks.org/the-stanford-literary-labs-narrative/.

51 See Ted Underwood, "A Genealogy of Distant Reading", *Digital Humanities Quarterly* 11, no. 2 (2017): § 7–11 and Thomas Weitin, Thomas Gilli and Nico Kunkel, "Auslegen und Ausrechnen. Zum Verhältnis hermeneutischer und quantitativer Verfahren in den Literaturwissenschaften", *Zeitschrift für Literaturwissenschaft und Linguistik* 46 (2016): 104. Peer Trilcke and Frank Fischer point out that Moretti, in his recent publications, no longer uses the term 'distant reading'. Instead, he would refer to 'computational criticism' or 'digital humanities'. See Peer Trilcke, and Frank Fischer, "Fernlesen mit Foucault? Überlegungen zur Praxis des *distant reading* und zur Operationalisierung von Foucaults Diskursanalyse", *Le foucaldien* 2, no. 1 (2016): 13. http://doi.org/10.16995/lefou.15.

52 See Lamping, Dieter."'Distant Reading', aus der Nähe betrachtet: Zu Franco Morettis überschätzter Aufsatzsammlung". *Literaturkritik.de*, no. 9 (2016). https://literaturkritik.de/id/22506#biblio.

53 See Kathryn Schulz, "Distant Reading: To Uncover the True Nature of Literature, a Scholar Says, Don't Read the Books", *New York Times*, June 26, 2011, B14.

54 Stanley Fish, "Mind Your p's and b's: The Digital Humanities and Interpretation", *New York Times*, last modified January 23, 2012, https://opinionator.blogs.nytimes.com/2012/01/23/mind-your-ps-and-bs-the-digital-humanities-and-interpretation/?_r=0.

55 Cf. Fotis Jannidis,"Digitale Geisteswissenschaften: Offene Fragen – schöne Aussichten", *Zeitschrift für Medien- und Kulturforschung* 10, no. 1 (2019): 63–70.

56 Martin, Mueller, "Digital Shakespeare, or towards a Literary Informatics", *Shakespeare* 4, no. 3 (2008): 290.

> which puts phenomena into the ghetto of an acronym that makes its practitioners feel good about themselves but allows the rest of the humanities to ignore them."[57]

Drawing on experiences from Google Earth and its possibility to seamlessly zoom in and out, Mueller instead proposes the term "Scalable Reading as a happy synthesis of 'close' and 'distant reading'".[58] For the study of literature, he suggests, in other words, mixing qualitative and quantitative methods, and thus combining the benefits of different methodological approaches. In contrast to distant reading, scalable reading promises infinite scaling which forms a connection from individual texts to huge text corpora, and also allows a change of scale, i.e., the methodological leap from qualitative interpretation to quantitative analysis—and vice versa.[59] This change of scale, however, is in no way as seamless as in Google Earth.[60] Rather, it requires well reflected operationalization to bridge the methodological gap.

Although the promise of infinite scalability is problematic, we argue that a reflected combination of qualitative and quantitative methods for the analysis of literature, as Mueller intends with his term 'scalable reading', can make for highly instructive and hypothesis-driven research; as numbers are in need of interpretation: they demand and "enable new and powerful ways of shuttling between 'text' and 'context'".[61] Hereafter, we will give three cursory examples of QuaDramA research that demonstrate the potential of a mixed-methods setup—or, in Mueller's words, the potential of 'scalable reading'—not only for bigger corpus analyses, but also for the understanding and interpretation of single literary texts. The examples try to show how qualitative and quantitative methods can build upon each other.

Our first example concentrates on Heinrich von Kleist's play *Die Familie Schroffenstein* (1803) and showcases a hypothesis-driven approach. With the establishment of the bourgeois family, authors analytically took gender differences into account for their writings, which are, e.g., important for the bourgeois tragedy.[62] Kleist's letters,

---

57 Martin, Mueller, "Scalable Reading", last modified April 26, 2020, https://web.archive.org/web/20211201185120/https://sites.northwestern.edu/scalablereading/2020/04/26/scalable-reading/.

58 Ibid.

59 Benjamin Krautter, and Marcus Willand, "Close, distant, scalable. Skalierende Textpraktiken in der Literaturwissenschaft und den Digital Humanities", in: Carlos Spoerhase, Steffen Siegel and Nikolaus Wegmann (eds.) *Ästhetik der Skalierung* (Hamburg: Felix Meiner Verlag, 2020), 86–87.

60 Cf. Thomas Weitin, *Digitale Literaturgeschichte. Eine Versuchsreihe mit sieben Experimenten* (Berlin: J.B. Metzler, 2021), 116.

61 Mueller, "Scalable Reading".

62 See Sigrid Lange, *Die Utopie des Weiblichen im Drama Goethes, Schillers und Kleists* (Frankfurt a.M.: Peter Lang, 1993), 138 and Ulrike Vedder, "Biblische Muster und ihre Spielräume in Kleists Familien- und Geschlechterordnungen", *Kleist-Jahrbuch* (2018): 87.

especially the remarks in his letters to Marie von Kleist, suggest that for him, gender might have been one of the more important poetological categories.[63] This hypothesis finds additional weight when one looks at the final act of *Die Familie Schroffenstein*. When the loving couple Agnes and Ottokar—they come from two hostile houses of the same family—change their clothes in the light of the danger that their nearing fathers are posing, metaphorically they also change their gender. To verify the hypothesis through supplementary quantitative analysis, we employed five word fields (love, family, reason, religion, and war) to investigate the semantics of the characters' speech. The results show that the speech parts of the male and female characters do differ. The values for male and female characters show gender-specific patterns regarding the word fields.[64] Ottokar and Agnes, however, do not fit in these patterns shown by the other characters. Instead, their values seem to be somewhat mirrored: When one looks at the distributions of the word fields, one sees that Ottokar rather speaks like the female characters, Agnes rather like the male ones. Thus, the semantics of Agnes' and Ottokar's character speech correspond to the lovers' final external identification when they are changing their clothes. In the end, this leads to the tragic death of both of them.[65]

The second example derives from the annotation of coreference chains in German plays, in this case Friedrich Schiller's *Die Räuber* (1781): When one looks at the chains of Franz Moor, he is brother and antagonist of Karl Moor, the head of a band of robbers, instructive findings can be made. Compared to the other main characters, Franz' utterances feature a lot of references to other characters.[66] It comes as no surprise, as his father and his brother are omnipresent in his thoughts. However, starting with the fifth act of the play, in which Franz will commit suicide, these references diminish, and his utterances become more self-centerd. As soon as father, brother and the main female character Amalia disappear as references, i.e., Franz' thought process no longer targets the perceived injustice of nature that he projected onto other characters, his own horrible intrigues collapse.[67]

While the first two examples showcase the use of qualitative and quantitative methods for the analysis and interpretation of single literary texts, our third exam-

63 See Marcus Willand and Nils Reiter, "Geschlecht und Gattung. Digitale Analysen von Kleists Familie Schroffenstein", *Kleist-Jahrbuch* (2017): 188, https://doi.org/10.1007/978-3-476-04516-4_16.

64 See Ibid., 187.

65 See Ibid., 189–190.

66 See Benjamin Krautter and Marcus Willand, "Vermessene Figuren – Karl und Franz Moor im quantitativen Vergleich", in: Peter-André Alt and Stefanie Hundehege (eds.): *Schillers Feste der Rhetorik* (Berlin, Boston: De Gruyter, 2021), 107–138.

67 See Ibid.

ple is geared towards a bigger corpus of dramas.[68] Based on literary history and theory, we operationalize dramatic character types in German-language drama. Applying Moretti's original idea of 'distant reading', i.e., the extensive study of secondary literature, we identified characters that belonged to one of the three types: 'schemer', 'virtuous daughter' and 'tender father'. Furthermore, we annotated those properties of the investigated characters that were referenced in the secondary literature. In total, we created a data set of 257 characters and their annotated properties.

These annotations are in turn used for the automatic classification of character types. The experiments show that the selected character types emerge as definable and automatically identifiable subsets within a bigger population. These results lay the groundwork for further, more extensive literary historical analyses of dramatic character types: When do character types emerge, when do they disappear? How do different character types interact within the plays?

All three examples combine quantitative and qualitative approaches for the analysis of dramatic texts. They reveal that the computational counting of surface text elements, such as character mentions or word frequencies, can provide insights into the meaning of a text, especially when it comes to counting the elements that mostly remain invisible to human readers. The examples are also to be understood as a statement that interdisciplinary mixed-methods approaches in digital humanities can fruitfully expand existing research discussions on different humanities disciplines such as literary studies.

Conclusions

In this article, we have argued for a consciously designed collaboration in mixed-methods projects. Interdisciplinary collaboration does not happen by itself, in particular in the digital humanities, where the underlying research assumptions, workflows, and interests can be quite different between D and H. In our project, the joint annotation of humanities concepts, technical workflows that ensure independence of researchers and a close integration of quantitative and qualitative findings are the core principles of successful collaboration.

As discussed in the introduction, we regard operationalization as one of the key challenges not only for our collaboration, but also for digital humanities in general. From our perspective, *good* operationalization of a literary studies concept is one that is both technically feasible and, at the same time, worth arguing about within

68 See Benjamin Krautter et al., '"[E]in Vater, dächte ich, ist doch immer ein Vater". Figurentypen und ihre Operationalisierung", in *Zeitschrift für digitale Geisteswissenschaften* (2020), http://dx.doi.org/10.17175/2020_007.

the literary studies community: while it may not be consensual, the literary studies community accepts it (and its results) as a contribution to the discussions in the field.[69] This is hard to achieve, as operationalization requires a lot of conceptual work depending on the degree of abstraction of the target categories. While measuring certain properties of a text is straightforward, e.g., how many words a character utters in a play, connecting these values to established concepts, thereby fostering the understanding of a specific text, the understanding of a writer's *œuvre*, or literary historical relations, requires sound theoretical reflection. Thus, filling the gap between 'concepts and measurement'[70] and then between measurement and meaning is one of the major challenges a mixed-methods project must tackle.

Digital humanities practitioners often refer to their discipline as a 'big tent' allowing for a multitude of methods, questions, objects, and approaches.[71] The disciplinary status together with its relation to the variety of 'origin disciplines' in the humanities is a subject of much debate. The digital humanities are an invaluable community that unites researchers employing digital, quantitative, and algorithmic methods to research questions from the humanities. The exploration of their potential should not be stopped solely because of disciplinary boundaries or because these methods still live in the shadows of many humanities disciplines. While development and discussion of digital methods are well accommodated in the digital humanities, once they are established, we envisage bringing the results of applying these methods back into the general field of the humanities. In our opinion, mixed-methods projects that focus and reflect on their collaboration in the light of their research questions and used methods are in a prime position to do so.

Bibliography

Adelmann, Benedikt, Melanie Andresen, Wolfgang Menzel, and Heike Zinsmeister. "Evaluating Part-of-Speech and Morphological Tagging for Humanities' Interpretation". In *Proceedings of the Second Workshop on Corpus-Based Research in the Humanities*, edited by Andrew U. Frank, Christine Ivanovic, Francesco Mambrini, Marco Passarotti, and Caroline Sporleder, 5–14. Wien, 2018. https://www.oeaw.ac.at/fileadmin/subsites/academiaecorpora/PDF/CRH2.pdf.

Can, Fazli, Tansel Özyer, and Faruk Polat (eds.). *State of the Art Applications of Social Network Analysis*. Cham: Springer, 2014.

69 The German word *satisfaktionsfähig* describes our intention well: The term originates in the context of duels and describes dualists as worthy of a duel.

70 Franco Moretti, '"Operationalizing": Or, the Function of Measurement in Modern Literary Theory,' *Pamphlets of the Stanford Literary Lab* 6 (2013): 1.

71 Cf. Patrik Svensson, "Beyond the Big Tent", in: Matthew K. Gold (ed.): *Debates in the Digital Humanities* (Minneapolis, London: University of Minnesota Press, 2012), 36–49.

Cohen, Margaret. "Narratology in the Archive of Literature". In *Representations* 108, 1 (2009): 51–75.

Cohen, Margaret. *The Sentimental Education of the Novel*. Princeton: Princeton University Press, 1999.

Eder, Maciej, Jan Rybicki, and Mike Kestemont. "Stylometry with R: A Package for Computational Text Analysis". *R Journal* 8, 1 (2016): 107–121. https://doi.org/10.32614/RJ-2016-007.

Ehrlicher, Hanno, Jörg Lehmann, Nils Reiter, and Marcus Willand. "Gattungspoetik in quantitativer Sicht. Das Werk Calderón de la Barcas und die zeitgenössischen Dramenpoetiken". In *LitLab Pamphlet* 9 (2020): 1–29. https://www.digitalhumanitiescooperation.de/wp-content/uploads/2020/04/p09_ehrlicher_lehmann_al_de.pdf.

Fischer, Frank, Mathias Göbel, Dario Kampkaspar, Christopher Kittel, and Peer Trilcke. "Network Dynamics, Plot Analysis. Approaching the Progressive Structuration of Literary Texts". In *Digital Humanities 2017. Conference Abstracts*, 437–441. Montréal, 2017. https://dh2017.adho.org/abstracts/DH2017-abstracts.pdf.

Fischer, Frank, Ingo Börner, Mathias Göbel, Angelika, Hechtl, Christopher, Kittel, Carsten Milling, and Peer Trilcke. "Programmable Corpora. Die digitale Literaturwissenschaft zwischen Forschung und Infrastruktur am Beispiel von DraCor". In *Abstracts of DHd*, edited by Patrick Sahle, 194–197. Frankfurt, Mainz, 2019. https://doi.org/10.5281/zenodo.2596095.

Fish, Stanley. "Mind Your p's and b's: The Digital Humanities and Interpretation". In *New York Times*. Last modified January 23, 2012. https://opinionator.blogs.nytimes.com/2012/01/23/mind-your-ps-and-bs-the-digital-humanities-and-interpretation/?_r=0.

Gius, Evelyn, and Janina Jacke. "The Hermeneutic Profit of Annotation: On Preventing and Fostering Disagreement in Literary Analysis". In *International Journal of Humanities and Arts Computing* 11, 2 (2017): 233–254. https://doi.org/10.3366/ijhac.2017.0194.

Hovy, Eduard, and Julia Lavid. "Towards a Science of Corpus Annotation: A New Methodological Challenge for Corpus Linguistics". In *International Journal of Translation Studies* 22, 1 (2010): 13–36.

Jannidis, Fotis. "Digitale Geisteswissenschaften: Offene Fragen – schöne Aussichten". In *Zeitschrift für Medien- und Kulturforschung* 10, 1 (2019): 63–70.

Jockers, Matthew L. *Macroanalysis. Digital Methods and Literary History*. Urbana, Chicago, Springfield: University of Illinois Press, 2013.

Krautter, Benjamin, Janis Pagel, Nils Reiter, and Marcus Willand. "Titelhelden und Protagonisten – interpretierbare Figurenklassifikation in deutschsprachigen

Dramen". In *LitLab Pamphlets* 7 (2018): 1–56. https://www.digitalhumanitiescooperation.de/wp-content/uploads/2018/12/p07_krautter_et_al.pdf.

Krautter, Benjamin, Janis Pagel, Nils Reiter, and Marcus Willand. "'[E]in Vater, dächte ich, ist doch immer ein Vater.' Figurentypen und ihre Operationalisierung". In *Zeitschrift für digitale Geisteswissenschaften* (2020). http://dx.doi.org/10.17175/2020_007.

Krautter, Benjamin, and Marcus Willand. "Close, distant, scalable. Skalierende Textpraktiken in der Literaturwissenschaft und den Digital Humanities". In *Ästhetik der Skalierung*, edited by Carlos Spoerhase, Steffen Siegel, and Nikolaus Wegmann, 77–97. Hamburg: Felix Meiner Verlag, 2020.

Krautter, Benjamin, and Marcus Willand. "Vermessene Figuren – Karl und Franz Moor im quantitativen Vergleich". In *Schillers Feste der Rhetorik*, edited by Peter-André Alt, and Stefanie Hundehege, 107–138. Berlin, Boston: De Gruyter, 2021.

Lamping, Dieter. "'Distant Reading', aus der Nähe betrachtet: Zu Franco Morettis überschätzter Aufsatzsammlung". In *Literaturkritik.de*, no. 9 (2016). https://literaturkritik.de/id/22506#biblio.

Lange, Sigrid. *Die Utopie des Weiblichen im Drama Goethes, Schillers und Kleists*. Frankfurt a.M.: Peter Lang, 1993.

Marcus, Solomon. *Mathematische Poetik*. Translated by Edith Mândroiu. Frankfurt a.M.: Athenäum Verlag, 1973.

Meister, Jan Christoph."Computerphilologie vs. 'Digital Text Studies'". In *Literatur und Digitalisierung*, edited by Christine Grond-Rigler and Wolfgang Straub, 267–296. Berlin, Boston: De Gruyter, 2013.

Moretti, Franco. "Conjectures on World Literature". In *New Left Review* 1 (2000): 54–68.

Moretti, Franco. "Graphs, Maps, Trees: Abstract Models for Literary History". In *New Left Review* 24 (2003): 67–93.

Moretti, Franco. *Graphs, Maps, Trees: Abstract Models for Literary History*. London and New York: Verso, 2005.

Moretti, Franco. "'Operationalizing': Or, the Function of Measurement in Modern Literary Theory". In *Pamphlets of the Stanford Literary Lab* 6 (2013): 1–13.

Mueller, Martin. "Digital Shakespeare, or towards a Literary Informatics". *Shakespeare* 4, 3 (2008): 284–301. https://doi org/10 1080/17450910802295179

Mueller, Martin. "Scalable Reading". Last modified April 26, 2020. https://web.archive.org/web/20211201185120/https://sites.northwestern.edu/scalablereading/2020/04/26/scalable-reading/.

Pagel, Janis, and Nils Reiter. "GerDraCor-Coref: A Coreference Corpus for Dramatic Texts in German". In *Proceedings of the 12th Language Resources and Evaluation Conference (LREC)*, 55–64. Marseille, 2020. https://www.aclweb.org/anthology/2020.lrec-1.7.pdf

Pagel, Janis, Nils Reiter, Ina Rösiger, and Sarah Schulz. "Annotation als flexibel einsetzbare Methode". In *Reflektierte Algorithmische Textanalyse*, edited by Nils Reiter, Axel Pichler, and Jonas Kuhn, 125–141. Berlin, Boston: De Gruyter, 2020.

Pfister, Manfred. *The Theory and Analysis of Drama*. Cambridge et al.: Cambridge University Press, 1988.

Pichler, Axel, and Nils Reiter. "Reflektierte Textanalyse". In *Reflektierte Algorithmische Textanalyse*, edited by Nils Reiter, Axel Pichler, and Jonas Kuhn, 43–59. Berlin, Boston: De Gruyter, 2020.

Pichler, Axel, and Nils Reiter. "Zur Operationalisierung literaturwissenschaftlicher Begriffe in der algorithmischen Textanalyse: Eine Annäherung über Norbert Altenhofers hermeneutischer Modellinterpretation von Kleists *Das Erdbeben in Chili*". In *Journal of Literary Theory* 15, 1–2 (2021): 1–29.

Pustejovsky, James, and Amber Stubbs. *Natural Language Annotation for Machine Learning: A Guide to Corpus-Building for Applications*. Sebastopol, Boston, Farnham: O'Reilly Media, 2012.

Reiter, Nils. Discovering Structural Similarities in Narrative Texts using Event Alignment Algorithms. PhD diss., Heidelberg University, 2014. https://doi.org/10.11588/heidok.00017042.

Reiter, Nils. "CorefAnnotator – a New Annotation Tool for Entity References". In *Abstracts of EADH: Data in the Digital Humanities*. Galway, 2018. https://doi.org/10.18419/opus-10144.

Reiter, Nils, Marcus Willand, and Evelyn Gius. "A Shared Task for the Digital Humanities Chapter 1: Introduction to Annotation, Narrative Levels and Shared Tasks". In *Cultural Analytics* (2019). https://culturalanalytics.org/article/11192.

Rösiger, Ina, Sarah Schulz, and Nils Reiter. "Towards Coreference for Literary Text: Analyzing Domain-Specific Phenomena". In *Proceedings of the Second Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, 129–38. Santa Fe, 2018. http://aclweb.org/anthology/W18-4515.

Schulz, Kathryn. "Distant Reading: To Uncover the True Nature of Literature, a Scholar Says, Don't Read the Books". *New York Times*, June 26, 2011, B14.

Svensson, Patrick. "Beyond the Big Tent". In *Debates in the Digital Humanities*, edited by Matthew K. Gold, 36–49. Minneapolis, London: University of Minnesota Press, 2012.

Trilcke, Peer, and Frank Fischer. "Fernlesen mit Foucault? Überlegungen zur Praxis des *distantreading* und zur Operationalisierung von Foucaults Diskursanalyse". In *Le foucaldien* 2, 1 (2016): 1–18. http://doi.org/10.16995/lefou.15.

Underwood, Ted. "A Genealogy of Distant Reading". In *Digital Humanities Quarterly* 11, 2 (2017): § 1–44. http://www.digitalhumanities.org/dhq/vol/11/2/000317/000317.html.

Underwood, Ted. "The Stanford Literary Lab's Story". Last modified February 11, 2017. https://www.publicbooks.org/the-stanford-literary-labs-narrative/.

Vedder, Ulrike. "Biblische Muster und ihre Spielräume in Kleists Familien- und Geschlechterordnungen". *Kleist-Jahrbuch* (2018): 87–99.

Weitin, Thomas. *Digitale Literaturgeschichte. Eine Versuchsreihe mit sieben Experimenten*. Berlin: J.B. Metzler, 2021.

Weitin, Thomas, Thomas Gilli, and Nico Kunkel. "Auslegen und Ausrechnen. Zum Verhältnis hermeneutischer und quantitativer Verfahren in den Literaturwissenschaften". In *Zeitschrift für Literaturwissenschaft und Linguistik* 46 (2016): 103–115.

Willand, Marcus, Benjamin Krautter, Janis Pagel, and Nils Reiter. "Passive Präsenz tragischer Hauptfiguren im Drama". In *Abstracts of DHd*, edited by Christof Schöch, 177–181. Paderborn, 2020. https://zenodo.org/record/3666690.

Willand, Marcus, and Nils Reiter. "Geschlecht und Gattung. Digitale Analysen von Kleists Familie Schroffenstein". In *Kleist-Jahrbuch* (2017): 177–95. https://doi.org/10.1007/978-3-476-04516-4_16.

Zweig, Katharina A. *Network Analysis Literacy. A Practical Approach to the Analysis of Networks*. Wien: Springer, 2016.

#HUMAN-IN-THE-LOOP

#Human-in-the-Loop is a neologism and not (yet) recorded in the Oxford English Dictionary. The OED, however, lists a specific meaning of loop in terms of computing as a "sequence of instructions which is executed repeatedly (usually with an operand that changes in each cycle) until some previously specified criterion is satisfied." * Against this background, the term suggests the intentional inclusion of human action into an automated process.

The inclusion is a crucial goal in the digital humanities (QuaDramA). Yet, the use within the mixed methods projects varies significantly. As far as they discuss it, the projects understand the term to encompass single or multiple interactions between (automated work of) algorithms and human actors, largely understood as humanities scholars or even as a general audience but not as scholars from computer science. As such, the addressed human acts outside the inner workings of computer science.

Beyond that, the projects set different foci. Some understand the said human to represent the target audience of a development such as, for example, the creation of an interface in the known relationship between client and contractor. Others perceive them as an indispensable conceptual partner of a joint endeavour in digital humanities.

In consequence, the understanding of the #Human-in-the-Loop oscillates primarily between three corner points: First, the (humanities) user as the target of a communication that originates in the realms and concepts of computer science (QuaDramA). Second, the processes and means to create an interface in which the looped-in actors ensure comprehension and functionality (Handwriting). Third, a communication platform-cum-monitoring mechanism between the epistemic systems of humanities and computer science aiming at hybridization of methods and concepts (DhiMu; ArchiMediaL). Beyond that, there is an additional corner point that touches the meaning of #Human-in-the-Loop: the human as a "means to generate the data base needed for the computer to learn" (ArchiMediaL), as the starting point of the loop(s).

The cloud of attributed meanings shows the neologism #Human-in-the-Loop as a crucial tool for negotiating the role of humans/humanities for digital humanities and/or the intersection of the fields. It will be interesting to see which components of this field of meaning remain permanently and whether new, more specific or general terminology will ultimately develop for some areas. Thus, the term can serve as an indicator of the development of interpretative powers within digital humanities.

* "loop, n.1.". in: *Oxford English Dictionary (OED)*, first published 1903; most recently modified version published online March 2022 with draft additions from 1997 and 2018, https://www.oed.com/ [accessed: 20.05.2022].

Title: Dhimmis and Muslims – Analyzing Multi-Religious Spaces in the Medieval Muslim World
Team: Max Franke, Steffen Koch (Stuttgart); Ralph Barczok, Dorothea Weltecke (Frankfurt)
Corpus: Research literature, medieval sources in various languages (Latin, Syriac, Arabic, Armenian etc.)
Field of Study: Medieval History, Oriental Studies, Information Visualization, Visual Analytics, Digital Humanities
Institution: Institut fürVisualisierung und InteraktiveSysteme (VIS) (University of Stuttgart), Department of History (Goethe University Frankfurt)
Methods: Historical text criticism, visualization and computer graphics research
Tools: Web-based data entry interfaces, web-based multiple-coordinated-views visualization of different aspects of the data
Technology: Python, JavaScript/TypeScript, D3.js, relational databases (PostgreSQL, MySQL)

Dhimmis and Muslims – Analyzing Multi-Religious Spaces in the Medieval Muslim World (DhiMu)

Ralph Barczok, Max Franke, Steffen Koch, Dorothea Weltecke

Abstract *For centuries, specific groups of non-Muslims living in regions under Muslim rule were tolerated and, in turn, forced to accept a lower legal status called the Dhimmi status. This legal construction and political pragmatism resulted in great religious diversity in medieval Muslim cities. However, the bias of the medieval writers, who faded out the existence of the other religious communities and the traditional historical narratives have so far inhibited an empirically grounded view of the diachronic and synchronic complexity of these multi-religious populations. This project aims to collect information from primary and secondary sources on religious groups in the urban centres of the Middle East. This information must be stored digitally to aid historians in conducting interactive visual analyses. Historians study the data of religious constellations and compare regions and time periods. By recording the level of confidence together with the data, they can consider the uncertainty of data during visual analysis to improve their judgment of collected historical evidence.*

Project description

The question of how to organize religious diversity in our societies is oft-discussed and politically relevant. At the same time, religious minorities in the Middle East are facing a deadly threat. Regular violence, civil war and religious cleansing will soon wipe out the last vestiges of the region's multi-religious past. The conflict between Sunnis and Shiites is accelerating both the process of topographical separation in cities, such as Baghdad, and the destruction of written and archaeological sources.

Today, a mono-religious situation has emerged in the Middle East. During the medieval centuries, however, the majority Muslim population tolerated specific groups of non-Muslims (especially Jews, Christians and Zoroastrians), but forced them to accept a lower legal status called the Dhimmi-status. This legal construction and political pragmatism led to a great religious diversity in the medieval Islamicate world, that is, the regions of Muslim rule. Thus, diverse Muslim strands (e.g., Sunnis, Shiites) as well as different churches of Eastern Christianity (e.g., the Melkite,

the Syriac Orthodox or the Church of the East) and Jewish traditions (Karaits, Rabbanites) lived side by side in the many cities. Often three or even four Christian bishops resided in the same city. Together they formed the fabric of everyday life and culture in the Islamicate world. This grade of religious diversity is not found in the medieval cities of the Latin, Byzantine and Caucasian worlds which were dominated by Christian rule.[1]

Analyzing historical evidence previously relied on printed maps. However, interactive geovisualization offers a more flexible way of presenting, inspecting and analyzing such complex information. Interactive geovisualization also provides new aspects such as the confidence of sources and the reproducibility of findings. Therefore, cooperation between historians and visualization researchers seemed natural for this project.

The project covers the historical period between the emergence of Islam in the 7th century, and the end of many Christian centres brought about by the invasion of the Mongol ruler Tamerlan at the end of the 14th century. Geographically, the project focuses on the areas of Muslim rule from North Africa to Western and Central Asia. This space is covered at the same time by the networks of the churches of the Eastern Christian patriarchates of Alexandria and Antioch, the diverse Eastern autocephalous churches, Jewish communities of different traditions, and also by other tolerated religious groups (Dhimmis) such as Zoroastrians. These tolerated religious groups are the focus of this project. Data on ephemeral religious movements, groups without an institutional framework and religious strands that Muslim authorities considered heretics or idolaters (and therefore did not tolerate), are not included in the project.

Scholars have understudied the diachronic and synchronic complexity of these multi-religious populations. Basic knowledge about the exact distribution and even the mere existence of religious groups in many places is still lacking. There are several reasons for this: Religious interaction in the medieval Muslim world, as such, is a new field that still requires a great deal of basic research (cf. JewsEast[2]; RelMin[3]). On the level of research, the dominant historical narratives of 'the Islamic culture' eclipse non-Muslim communities. The study of Eastern Christianity and of Jews in the Muslim world also predominantly focuses on the history of one denomination. While the relationship of Christians, Jews, Zoroastrians and others to their Muslim

1 Dorothea Weltecke, „Zum syrisch-orthodoxen Leben in der mittelalterlichen Stadt und zu den Hūddōyē (dem Nomokanon) des Bar ʿEbrōyō", in: Peter Bruns and Heinz-Otto Luthe (eds.): *Orientalia Christiana. Festschrift für Hubert Kaufhold zum 70. Geburtstag* (Wiesbaden: Harrassowitz Verlag, 2013).

2 Alexandra Cuffel, "Jews and Christians in the East: Strategies of Interaction from the Mediterranean to the Indian Ocean", https://www.jewseast.org/.

3 John Tolan, "RelMin: The Legal Status of Religious Minorities in the Euro-Mediterranean World (5th–15th Centuries)". https://www.cn-telma.fr/relmin/index/.

overlords is a traditional field of research, the analysis of concrete spatial constellations that integrate different groups is still a desideratum. Thus, historical narratives are still based on impressions of singular sources and traditional views rather than empirical data. The historical sources, however, suppress competing religious groups. Often, they primarily report on matters concerning their own community and are less interested in describing everyday interactions with others in any detail.

We want to help overcome the persistent confessional tendency of research and treat all religious communities as one population that shares a common social space. The historical and textual data that we have gathered from a variety of printed sources in different languages and qualities is being integrated into a single repository for the first time. To establish geographical coordinates, we use digital databases of projects like Pleiades[4] with subprojects such as DARE[5], syriaca.org[6], the Prosopographie der mittelbyzantinischen Zeit[7], the Encyclopaedia of Islam[8], Encyclopaedia Judaica[9] and many more. Digital visualization makes our historical data more approachable, surmounting the limits of linear text reading. This way, we seek to produce visual means to study hotspots of multi-religious existence and provide a diachronic as well as geographical perspective to the communities' expansion and contraction.

4 Roger Bagnall and Richard Talbert (eds.), "*Pleiades: A Gazetteer of Past Places*" *https://pleiades.stoa.org/*.

5 Johan Åhlfeldt, "DARE: *Digital Atlas of the Roman Empire*", https://dh.gu.se/dare/.

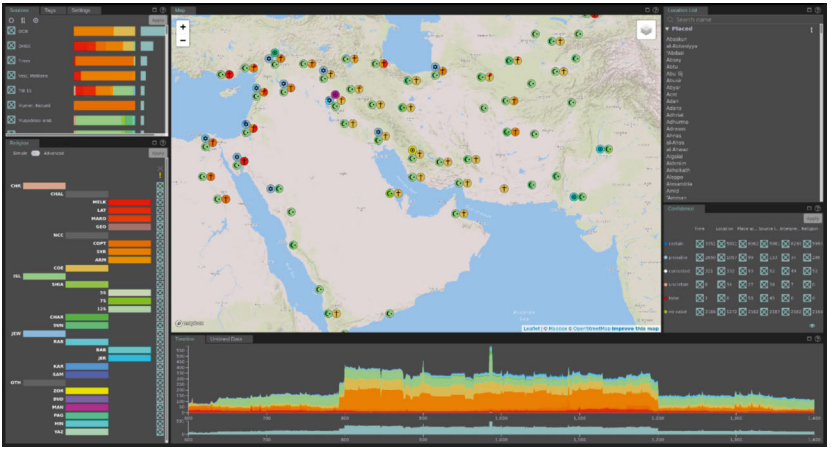
6 Nathan P. Gibson, David A. Michelson and Daniel L. Schwartz (eds.), "Syriaca.org" http://syriaca.org/.

7 Ralph-Johannes Lilie, et al. (eds.), *Prosopographie der mittelbyzantinischen Zeit. Erste Abteilung (641–867). Nach Vorarbeiten Friedhelm Winkelmanns erstellt von Ralph-Johannes Lilie, Claudia Ludwig, Thomas Pratsch, Ilse Rochow und Beate Zielke unter Mitarbeit von Wolfram Brandes und John Robert Martindale. 7 Volumes* (Berlin: de Gruyter, 1998–2002); Ralph-Johannes Lilie, et al. (eds.), *Prosopographie der mittelbyzantinischen Zeit. Zweite Abteilung (867–1025). Nach Vorarbeiten Friedhelm Winkelmanns erstellt von Ralph-Johannes Lilie, Claudia Ludwig, Thomas Pratsch und Beate Zielke unter Mitarbeit von Harald Bichlmeier, Bettina Krönung, Daniel Föller sowie Alexander Beihammer und Günter Prinzing. 9 Volumes* (Berlin: de Gruyter, 2009–2013).

8 Peri J. Bearman, et al. (eds.), *The Encyclopaedia of Islam. Second Edition* (Leiden: Brill, 1960–2004).

9 Michael Berenbaum and Fred Skolnik (eds.), *Encyclopaedia Judaica. Second Edition. 22 Volumes* (Detroit: Thomson Gale, 2007).

Fig. 1: A screenshot of the interactive prototype developed for the Dhimmis & Muslims project. The prototype visualizes historical evidence data on the known presence of different institutionalized religious denominations in cities of the medieval Middle East. Different aspects of the data are visualized in multiple coordinated views, meaning that the application of a filter in one view affects the data in all other views as well. Also, the selection of data in one view (brushing) will highlight the same data in all other views (linking). This allows fine-grained analyses and the orthogonal combination of filters on different aspects of the data, together with an understanding of the relationships between different aspects of the data, such as temporal extent and geographical extent. The map view in the middle shows glyphs representing cities or, at lower zoom levels, clusters of cities. A glyph consists of up to four parts, three representing the monotheistic religions of Christianity, Islam, and Judaism, and the fourth representing other religions. The glyphs in combination with the interactive filtering allow exploration of the coexistence of different religions, the main focus of the project.



The results are evaluated on their confidence related to time, place, source and interpretation. The information is accessible through an interactive multiple-coordinated-views visualization (see Figure 1) process, which allows users to aggregate, correlate and compare the data visually in geo-temporal contexts while they extract the information. This approach allows the scholar to develop grounded hypotheses about the multi-religious life in Asia and North Africa. Our visualization approach is implemented by using web technology, which allows the scholar to access the data and visualization from their workplace by using their familiar web browser, instead of going through complicated installation and configuration processes of proprietary software.

The information from primary and secondary sources is stored in a relational database and is accessible for interactive analysis in the visual interface. In addition,

the final approach will offer interactive access to the annotated texts in accordance with their respective copyright restrictions. If the text is not publicly available, either a hyperlink or at least a bibliographical note is given.

Method reflection

Preliminary remarks

Our goal is to advance research and methods for both cooperative partners: We want to revive historiographical methods of heuristics and source analysis and simultaneously strengthen methods of visual analytics, advance the understanding of visual communication of data, interaction design and human cognition. Thus, we are not interested in one-sided cooperation, where information researchers simply design a tool or an environment according to the historians' suggestions. Since the beginning of our cooperation, we have taken pains to avoid such a dynamic. Visualization researchers are prepared to assist humanities scholars in deploying established techniques for their specific research questions. As the complexity of the joint project surpasses the means of such a simple implication, our cooperation must integrate the workflow and creative freedom of both sides. We carried out requirement analysis before developing concrete solutions and created a control loop to evaluate the implemented techniques. The visualization researchers also experimented with new, unproven visualization techniques and suggested solutions that the historians had not previously envisaged. We collaborate closely and discuss our problems, expectations, ideas and concerns on a regular basis. In this way, we ensure that both teams achieve their mutual goal. The structure of our contribution reflects our dialogical method and will represent the differences and similarities in our perspectives on the topics raised for a more general evaluation of our approach to a combination of historical (H) and visualization (VIS) research.

A. Quantification: How to quantify a research question from a domain field to make quantitative research possible? What is the status of statistics involved in your project?

H: For the historians, there are two basic quantitative questions at the core of this project: How many cities in the Islamic world hosted institutionalized communities of Dhimmis? How many different religious communities were living in these cities at a given point in time? Based on our previous research, we started out with a sample of 250 cities with up to 5 different communities in each. Our continuing research into written text sources led to approximately 450 cities and towns with up to 10–15 communities. These numbers represent an important initial result. While

historians are aware of the fact that the Muslim world and Muslim cities were often multi-religious, the exact distribution of non-Muslim communities is unknown.[10] We currently lack even basic knowledge about the existence of religious groups in many places. Historical narratives are largely based on impressions rather than hard data. Research is hampered by the fact that the written sources of the Middle Ages tend to mask other religious communities. Thus, an integrated and empirically valid result of the research question is only possible with the visualization of data accumulated from different religious strands. Obviously, these two quantitative questions require only basic mathematics. The data do not allow of sophisticated statistical research. Additionally, the data and the number of texts we work with are rather small compared to other projects on which our partners usually cooperate.

VIS: From a visualization research point of view, collaboration with humanities scholars presents both an opportunity and a challenge: the former, because field-specific problems as well as different perspectives and workflows can result in novel and interesting visualization approaches being developed; the latter, because it is difficult to evaluate such approaches with user studies due to the relatively few experts that are typically involved. This limits the possibilities of a quantitative evaluation that would require a larger number of test subjects, ideally not involved in the project. Studies in such collaborative projects are thus often reduced to qualitative evaluations, thus limiting insights into the generalizability of the developed visualization approaches. In other words, for visualization research itself, these projects do not necessarily yield valuable new research contributions.

B. Qualitative data: How to deal with qualitative data/insights?

H: Due to the nature of our sources and the time frame of our project, we had to limit our inquiry to established religious communities. But which religious communities do we consider to be 'established', especially in their formative period from the seventh to the ninth century? At which point in time do we consider them to be 'established' communities? Additionally, the question of what we consider to be a city or a town also arises. On which definition should we base our decision regarding which settlements to study?[11] How do we deal with the fact that villages

10 Janet L. Abu-Lughod, "The Islamic City – Historic Myth, Islamic Essence, and Contemporary Relevance" *International Journal of Middle East Studies* 19 (1987): 155–76; Gulia Neglia, "Some Historiographical Notes on the Islamic City with Particular Reference to the Visual Representation of the Built City", in: Salma Khadra Jayyusi, Renata Holod, Attilio Petruccioli and André Raymond (eds.), *The City in the Islamic World (HdO.Abt. 1 v. 94)* (Leiden, Boston: Brill, 2008), 1–46.

11 Abu-Lughod, "The Islamic City"; Paul Wheatley, The Places Where Men Pray Together. Cities in Islamic Lands, Seventh through the Tenth Centuries (Chicago: University of Chicago Press, 2001); Edmund Bosworth, Historic Cities of the Islamic World (Boston: Brill, 2007).

grew into cities over time, while other cities sank into oblivion? Here, the historians had to make many hard decisions and reinterpret and readjust our definitions in view of particular groups and settlements in time.

VIS: From a data-centric viewpoint, historical data are very much qualitative data, as very few aspects of them can be quantified. Many of the sources including text, images, etc., are digital representations of artefacts. From a computer science perspective, such historical and hermeneutical information lacks structure, making it difficult to handle it computationally without interpretation and assessment by historians and field-adequate preprocessing (see paragraph D). This can greatly increase computational complexity and complicates algorithmic summarization as well as reasoning while working with this data.

The same can be said about evaluative feedback collected from field experts to help in visualization research. As already discussed in A, these evaluations are mostly qualitative. While some statistics can be collected during longitudinal studies, most feedback is collected in the form of field experts' statements and as anecdotal findings.

C. Uncertainty: Historians and visualization researchers have different ideas of the concept of uncertainty.

VIS: In computer science, 'uncertainty' is an overloaded term. It is often used as a quantitative measure of data quality, for example, for the output of a machine learning algorithm. However, when historians use the term, they instead mean a qualitative measure, where the granularity and interpretation can vary individually, both between different points of data and different field experts. In the project, we decided to use the term 'confidence' to refer to these aspects. Quantifying and digitally storing such hermeneutical interpretations can be tricky: assessing them with numbers may indicate a measure of precision which is not there and, therefore, creates an additional source of uncertainty. We use a small number of categories to specify confidence, which, additionally, ensures a homogeneous interpretation by different researchers.[12]

H: For historians, confidence is a traditional mode of precise narrative description. In principle, however, confidence has been approached by visualization techniques established in the field of history, e.g., printed maps. Historians deal with different layers of uncertainty as temporal, geographical and text-critical confidence. With historical hermeneutical methods, the combination of very uncertain information may still produce absolutely trustworthy results as an historical assertion. On the

12 Max Franke, Ralph Barczok, Steffen Koch and Dorothea Weltecke, "Confidence as First-Class Attribute in Digital Humanities Data" in: Proceedings of the 4th VIS4DH Workshop. IEEE (2019).

other hand, there may be sources with a certain origin and date that nonetheless contain false or anachronistic information, as identified through historical source critique. Thus, the information they yield does not amount to an historical assertion of any validity.

D. Create interpretable models: Choosing a useful modelling framework.

VIS: In computer science, 'interpretable' means interpretable for the computer or the algorithm. This invariably means quantification, because computers, on a base level, can only handle numbers. On a more abstract level, this means that all data need to be conformed to some structure of values. Assumptions have to be made about invariants in the data. Almost always, the reality is more complex than the data model, and, consequently, the data must be modified to fit into the model. The goal of data visualization is to bridge the gap between data models that can be interpreted by computers and humans by finding suitable mapping to generate the visualization automatically. The data model must be created carefully so that the data are not oversimplified. After all, the resulting visualization can only represent the data as represented in the model. We have found that restricting data complexity can, to a degree, be alleviated by allowing free-text commentary to be attached to arbitrary data attributes. However, these are data that exist outside the framework of the model, such as a note pencilled in the margin of a book. As such, it is difficult to include such data in algorithmic processes and automatic visualization, and they only exist for the benefit of providing context to researchers in the humanities.
H: Historians use models to define and specify the object of their research. Preliminary research on the multi-religious history of the Middle Ages and on Dhimmis in the Muslim world shows that crowdsourced databases on Christian and Jewish communities in the Middle East generally do not yield useful results. They are simultaneously too large in scope and too arbitrary with respect to their data. Therefore, it was necessary to restrict the data we wanted to collect. A useful starting point in the spatial aspect was cities. This starting point reduced the number of places and restricted us to considering those places that were more likely to be covered by our sources. The main challenge was to single out those cities that were most relevant to our research question without distorting our results. We focused on adherents of established religions, since other, ephemeral forms of religious convictions proved to be too diffuse in their description and too unsteady to leave a social impact. Restriction to institutionalized groups rather than individuals was necessary to yield comparable and robust data that could not be found in information on the Dhimmi population in general.

From our own field of specialization, the history of Eastern Christianity, we knew that Christian history could be a good starting point to identify cities and

towns with Dhimmi groups.[13] Since the fourth century, Christian churches have appeared as institutionalized structures with a monarchical city bishop and their integration in the diocese of the patriarchs. As bishops were systematically recorded in lists, their spatial existence was much easier to retrieve than in the case of Jewish and Muslim communities. For Christians, unlike for Jews and Muslims, detailed mapping was already available. Cities with bishops thus formed the basis of our databases. In these cities, we used synagogues and mosques, judges and schools to determine which other religious strands established religious institutions. We added cities categorized by medieval Muslim geographers as central cities, and by Jewish tradition as major centres of learning, and investigated them accordingly.

E. What is the status of machine learning in your project? What do you mean by learning within a DH framework for the hermeneutical idea of learning in contrast to the current machine learning?

VIS: We do not use machine learning.

F. How to read the neural clustering/classification by the hermeneutical 'human in the loop'?

We are unfamiliar with the concept of 'neural clustering'. Therefore, our answer relates solely to the concept of 'human in the loop'.

VIS: In visual analytics, the human in the loop is a key figure for generating knowledge. The idea is that users not only consume visualized data but also interact with them, in an iterative way, to steer and improve data quality, data analysis and data visualization. This interaction can improve data quality and user trust in the visualized data, incrementally and on many levels.[14] Collaboration between field experts and visualization researchers may be essential for such incremental improvement: discussions between the two parties lead to the development of a visual analytics system tailor-made for the tasks the field experts want to perform and for their individual workflows. Field experts simultaneously learn about best practices and common techniques in VIS, which, in turn, increases their trust in the system they are using.

We believe that such iterative approaches match well with hermeneutical procedures, given that there are typical restrictions in terms of scope and expressiveness

13 Dorothea Weltecke, "60 years after Peter Kawerau. Remarks on the Social and Cultural History of Syriac-Orthodox Christians from the XIth to the XIIIth Century", *Le Muséon* 121 (2008): 311–35.

14 Dominik Sacha, Hansi Senaratne, Bom ChulKwon, Geoffrey Ellis and Daniel A. Keim, "The Role of Uncertainty, Awareness, and Trust in Visual Analytics", *IEEE Transactions on Visualization and Computer Graphics* 22 (2016): 240–249.

for algorithmic approaches. However, this does not stop scholars from leaving the traditional means.

G. (How) do you visualize the results of your dh-project? What is the status of the data visualization as a research tool/web interface?

H: Visualization was and is the main goal of our research. Traditionally, historians have used graphics and maps to visualize data, even in medieval chronicles. Graphics and maps are even more indispensable elements of data-driven research like ours. As existing digital methods would not answer our research questions, nor were existing analogous methods applicable to this case, the historians were keen to cooperate and develop new modes with greater complexity.
VIS: In this project, data visualization is not merely a means to present our results but a method for historians to reason about and create results by using an interactive visualization approach. The incremental design and evaluation of the visualization are intrinsic parts of our cooperation.

Technically, we provide a web-hosted, multiple-coordinated-views visualization for developing hypotheses and depicting results and findings. The temporal and spatial aspects of the data are visualized in separate views. Other views show the religious denominations and the uncertainty of other aspects of the data. To support a better understanding of the relationships between different aspects of the data, the views are all visually and interactively linked. Because of limits of screen size and human cognition, the data are initially aggregated; and additional detail as well as local nuances can be retrieved on demand.

Discussion

A. How did digitalization change your research/scale of research?

This question is apparently addressed only to the humanities scholars. For joint projects, one should also ask whether and how the contact with humanities and their sources changes computer science research.
H: Historians are most keen to benefit from the sheer enhancement and accessibility of data through digitalization.[15] Historical research on the period from 500 to 1500 welcomes digitization of sources and access to databases, collections, archives and museums. Medievalists are less active in applying digital techniques to gather

15 Laura Busse et al. (eds.), Clio-Guide. Ein Handbuch zu digitalen Ressourcen für die Geschichtswissenschaften. 2. erw. und aktualisierte Auflage (Berlin: Clio-online und Humboldt-Universität zu Berlin, 2018).

information and develop critical methods. The challenge of Semitic languages notwithstanding, European text recognition and geographical visualization are among the fields best developed in this area. This focus is easily explained with the importance of textual sources for medievalists and the spatial aspect of all historical developments (comp. The HGIS Club).[16] Thus, from the historian's perspective our research questions as such were designed within a fairly well-established field. Its focus on aspects of space and on the history of multi-religious constellations represents current trends in the history of religions and its cross-cultural approach. The project can benefit from further development of DH in the field as well (comp. Syriaca.org).[17]

At the same time, this project changes our research as we must constantly translate narratives into clear-cut definitions and numbers, and learn about professional modes of visualization. This 'culture shock' increases our awareness of the constructedness of visual representation and of research areas such as 'critical cartography', which ultimately fits in nicely with our background in post-colonial theory.

VIS: Digitalization is a prerequisite for all computational science disciplines, including visualization research.

B. What are the conflicts you experience in your own fields?

Possible conflicts may arise from researchers from different disciplines being unfamiliar with one another's terminology, their way of conducting research, and their general background knowledge of the respective fields. While terminology is discussed thoroughly when joint projects and proposals are planned, there seem to be reservations about such collaborations in both fields of computer science and humanities when it comes to publications. This may partially stem from a lack of knowledge, but there are other plausible causes as well.

VIS: The following reflects our observations made in the field of interactive data visualization in several mixed-method projects.

In general, the data visualization community is open to collaborative research endeavours for a good reason. While there is basic research in this discipline that can be undertaken without specific tasks or field-related requirements, data visualization would not make much sense without human interactants from other fields who perceive the visualized data and take action accordingly.

16 Kilian Schultes, Armin Volkman and Stefan Geißler, "The HGIS Club: Interdisziplinärer Arbeitskreis Historische Geographische Informationssysteme & Digital Humanities", http://hgis.club/.

17 Gibson, Michelson and Schwartz „Syriaca.org".

Nevertheless, visualization research, similar to all other research disciplines, aims at advancing research beyond the state of the art, meaning that published work ideally contains original and novel visualization developments. Such novelty, however, may not be a collaborator's primary goal in mixed-methods projects. Often, the application of well-known visualization approaches is exactly what is required to create suitable solutions. While successful approaches can be published in the form of an application paper or design study, they are considered less impactful than general improvements and are more difficult to publish in top visualization research venues. Publication methods and research interests also have an effect on the sustainability of solutions. As things stand, the development of novel approaches in the visualization community is often fast-paced and the sustainable realization and application of these concepts are considered to be implemented by those who would like to use these approaches. Long-term development in VA research profits from exchanging these approaches between fields, adapting them beyond individual project solutions, and combining as well as developing knowledge and experience over time.

One way to circumvent these problems is to collaborate on mixed-methods problems that present the opportunity to both pursue a valid goal in the humanities/social sciences and make progress in terms of visualization development. Obviously, this restricts possible joint project setups. However, the field of data visualization is becoming increasingly receptive to mixed-methods projects. Visualization venues are explicitly requesting submission of mixed-methods approaches and specific mixed-methods venues are being established (e.g., the VIS4DH, an event co-located with the most prominent visualization conference IEEE VIS, will be held for the seventh time this year (2023)). In addition, many funding agencies are becoming aware of the value of sustainable (visualization) developments and beginning to offer funding schemes.

H: As stated, this project can only count established Dhimmi communities and cities. In terms of the complexity of religious history, this is a very limited focus, which may even be considered misleading. We are asked to justify why and how these limits are necessary and still yield robust and methodologically sound historical results. As our approach privileges institutions and social structures rather than ephemeral agency, fluidity of demarcations and change, the design will attract criticism from religious and cultural studies, too: From this point of view, the project is methodologically conservative. Here, it will be necessary to explain why stable structures and labelled affiliations may not reflect the complexity of human religious interaction. But we will be able to point out some of its conditions not taken into account before.

C. Details versus abstraction: is this a conflict/problem of reduction/simplification?

VIS: No! In our understanding, abstraction, reduction and simplification are intrinsic characteristics of data visualization that help comprehend data from abstracted perspectives and explore large and complex data. We would even go so far as to claim that a lack of abstraction or reduction indicates that there is room for improvement in an interactive visualization approach. Having said that, it is, of course, beneficial or even necessary to support interactive ways that allow scholars to understand and verify the creation of visual data abstractions by showing the (digitized) data from which they were derived.

Apart from the optimization of analysis workflows, there are also other constraints that suggest simplification, such as limited resources. They manifest both in the limit to hardware resources, such as screen estate, and in the number of elements that human cognition can possibly process at one time. Even modern screens can hardly display as much information at one time as, for instance, a large, printed map, because of spatial and screen resolution limitations. However, computer-backed visualization can harness the computing power to change the visualization on demand, thus making it interactive. Visualization literature mostly agrees on adhering to the visual-information-seeking principle, introduced by Shneiderman: "Overview first, zoom and filter, then details-on-demand."[18] Following that mantra, interactive visualizations initially show aggregated data; and, on interaction, more details are shown for a smaller subset of that data.

In our discussions, we noticed that introduction of interactive workflows in the sense of Shneiderman's principle was a concept the historians first had to adjust to; they still very much desired to analyze—and generate—static images. The advantage of algorithmic visualization is the reproducibility of such static images, given the input data and interaction sequences. We found that the choice of aggregation was essential to adopting the resulting visualizations. In particular, the representation of aggregated data should be noticeable and meaningful; and the degree of aggregation should not be greater than what is dictated by the resource limitations. Further, we agreed that the level of aggregation or abstraction should be kept consistent at one point in time, even in locations with low information density. Doing otherwise yielded a false sense of high information density. Again, we found that a balance must be struck between the two disciplines and that design choices, such as the amount and method of aggregation, should be discussed and communicated clearly in the visualization.

18 Ben Shneiderman, "The Eyes Have It: A Task by Data Type Taxonomy for Information Visualizations", in: *Proceedings of the Symposium on Visual Languages. IEEE* (1996).

H: In terms of our project, three aspects of our initial model proved to be problematic for the historians:

1. The visualization may suggest quantitative relations according to the size of dots or aggregations. To prevent them from being misleading, it is necessary to adapt the modelling of the visualization and categories. As historians interpret large dots as having more importance and greater weight, for example, sizes have to be adapted to meet their traditional expectations of map symbols.
2. As religious distinctions only develop over time, it is difficult to determine the exact point at which an affiliation can be defined. Here, historical research risks falling prey to traditional narratives and assumptions. Our sources were often not specific enough in their designation of strands, so we needed to assign groups to a 'generic' category, e.g., 'generic Judaism'. In the database, this could signify that the group in question either existed before the segmentation of Judaism in the Middle East into Karaits and Rabbanitesor no specific information is extant, or that demarcations did not matter in this region. The visualization, on the other hand, conflates this 'generic' group with the sum of all the specific groups. In other words, the visualization does not differentiate between a settlement with only generic and vague information on Jews and another with a variety of well-defined Jewish groups (e.g., Rabbanites, Karaits and Samaritans). Thus, the individual database entries are more precise than the visualization. However, an aggregated view of the data, while not showing every detail at once, could reveal patterns and connections more clearly than a closer reading of the individual data entries would do.
3. We visualize the diversification of religious strands with hierarchical trees to indicate affiliation. But the tree graph seems to imply a model of historical partition and multiplication. The historical process is better described, however, as a reduction of the wide range of religious propositions and practices: Theological standards were constantly negotiated, as were decisions of orthodoxy and heterodoxy. Thus, the fluid beliefs and practices were reduced to more defined categories. New propositions and practices emerged and groups changed over time. We will have to explain these insights in a traditional historical narrative—or create additional graphs as illustrations.

D. Do you consider yourself a DH-laboratory?

Neither partner is exactly sure what the term 'DH-laboratory' implies. Our approach for temporal and geographical visualization will admit of a grounded hypothesis on the scale of multi-religious life in Asia and North Africa. On the one hand, we expect our project to bring the heuristics and analyses of historical information to a new level and create real progress in visualization research. Our approach and research

aims could not be realized with traditional historical techniques. We also hope to create a pilot-model for other complex research in historical topography. On the other hand, this joint project does not change theories of historical research or visual analytics. While we develop techniques of modelling and interpreting our specific data historically and with digital computation, the basic methods and assumptions underlying our cooperation, that is historical research and visual analytics, do not change.

Prospect for future DH research

VIS: DH is a very interesting direction for data visualization research because of the many tasks and research questions that are considerably different from those we see in collaborative projects with engineering, natural science and life science disciplines. In the light of these tasks and questions, we see great opportunities in joint projects with the humanities and social sciences. However, there are also differences, as elaborated in greater detail in B. One of the most important aspects of successful DH research, at least in terms of funding, seems to be sustainability of methods and tools. Interdisciplinary research is an exciting setup that can pave the way for DH research.

H: So far, the application of digital resources and techniques has changed neither the epistemological frame of our historical research nor the methods in general. Additionally, the technical expertise requires specific qualifications. This project proves that professional computer science researchers are the ideal partners for digital history problems. However, the more historians with digital skills or computer scientists with historical skills build the bridge between the two areas of research, the better basic training could be provided for all scholars. Traditionally, historians had to adapt their research designs to the conditions provided by the nature and number of sources. They always had to develop specific skills to use these sources, such as dating of charters, paper or artefacts. In a similar fashion, they will now approach digital sources and techniques with new skills and new methods of historical critique.

In this project, our mutual goal was to advance research and methods for both partners and provide new answers—as well as new questions—to both of our fields. We want to revive historiographical methods of heuristics and source analysis and, at the same time, strengthen the methods of visual analytics and information visualization. The challenges and results of our cooperation demonstrate the validity of this approach.

Bibliography

Abu-Lughod, Janet L. "The Islamic City—Historic Myth, Islamic Essence, and Contemporary Relevance". In International Journal of Middle East Studies 19 (1987): 155–76.

Åhlfeldt, Johan. "DARE: Digital Atlas of the Roman Empire". Accessed January 19, 2021. https://dh.gu.se/dare/.

Bagnall, Roger and Richard Talbert (eds.). "Pleiades: A Gazetteer of Past Places", https://pleiades.stoa.org/ [accessed: 19.01.2021].

Bearman, Peri J. et al. (eds.). The Encyclopaedia of Islam. Second Edition. Leiden: Brill, 1960–2004.

Berenbaum, Michael and Fred Skolnik (eds.). Encyclopaedia Judaica. Second Edition. 22 Volumes. Detroit: Thomson Gale, 2007.

Bosworth, Edmund. Historic Cities of the Islamic World. Boston: Brill, 2007.

Bruns, Peter and Heinz-Otto Luthe (eds.) Orientalia Christiana. Festschrift für Hubert Kaufhold zum 70. Geburtstag (Eichstätter Beiträge zum christlichen Orient 3). Wiesbaden: Harrassowitz Verlag, 2013.

Busse, Laura et al. (eds.). Clio-Guide. Ein Handbuch zu digitalen Ressourcen für die Geschichtswissenschaften. 2. erw. und aktualisierte Auflage. Berlin: Clio-online und Humboldt-Universität zu Berlin, 2018.

Cuffel, Alexandra. "Jews and Christians in the East: Strategies of Interaction from the Mediterranean to the Indian Ocean", https://www.jewseast.org/ [accessed: 19.01.2021].

Franke, Max, Ralph Barczok, Steffen Koch, and Dorothea Weltecke. „Confidence as First-Class Attribute in Digital Humanities Data". In Proceedings of the 4th VIS4DH Workshop. IEEE (2019).

Gibson, Nathan P., David A. Michelson, and Daniel L. Schwartz (eds.). "Syriaca.org", http://syriaca.org/ [accessed: 19.01.2021].

Lilie, Ralph-Johannes et al. (eds.). Prosopographie der mittelbyzantinischen Zeit. Erste Abteilung (641–867). Nach Vorarbeiten Friedhelm Winkelmanns erstellt von Ralph-Johannes Lilie, Claudia Ludwig, Thomas Pratsch, Ilse Rochow und Beate Zielke unter Mitarbeit von Wolfram Brandes und John Robert Martindale. 7 Volumes. Berlin: de Gruyter, 1998–2002.

Lilie, Ralph-Johannes et al. (eds.). Prosopographie der mittelbyzantinischen Zeit. Zweite Abteilung (867–1025). Nach Vorarbeiten Friedhelm Winkelmanns erstellt von Ralph-Johannes Lilie, Claudia Ludwig, Thomas Pratsch und Beate Zielke unter Mitarbeit von Harald Bichlmeier, Bettina Krönung, Daniel Föller sowie Alexander Beihammer und Günter Prinzing. 9 Volumes. Berlin: de Gruyter, 2009–2013.

Neglia, Giulia. "Some Historiographical Notes on the Islamic City with Particular Reference to the Visual Representation of the Built City". In The City in the Is-

lamic World (HdO.Abt. 1 v. 94), edited by Salma Khadra Jayyusi, Renata Holod, Attilio Petruccioli, and André Raymond, 1–46. Leiden, Boston: Brill, 2008.

Sacha, Dominik, Hansi Senaratne, BomChul Kwon, Geoffrey Ellis, and Daniel A. Keim. "The Role of Uncertainty, Awareness, and Trust in Visual Analytics". In IEEE Transactions on Visualization and Computer Graphics 22 (2016): 240–249.

Schultes, Kilian, Armin Volkman, and Stefan Geißler. „The HGIS Club: Interdisziplinärer Arbeitskreis Historische Geographische Informationssysteme & Digital Humanities", http://hgis.club/ [accessed: 19.01.2021].

Shneiderman, Ben. "The Eyes Have It: A Task by Data Type Taxonomy for Information Visualizations". In Proceedings of the Symposium on Visual Languages. IEEE (1996).

Tolan, John. "RelMin: The Legal Status of Religious Minorities in the Euro-Mediterranean World (5th–15th Centuries)", https://www.cn-telma.fr/relmin/index/ [accessed: 19.01.2021].

Weltecke, Dorothea. "60 years after Peter Kawerau. Remarks on the Social and Cultural History of Syriac-Orthodox Christians from the XIth to the XIIIth Century". In Le Muséon, 121 (2008): 311–35.

Weltecke, Dorothea. "Zum syrisch-orthodoxen Leben in der mittelalterlichen Stadt und zu den Hūddōyē (dem Nomokanon) des Bar ʿEbrōyō". In Orientalia Christiana. Festschrift für Hubert Kaufhold zum 70. Geburtstag (Eichstätter Beiträge zum christlichen Orient 3), edited by Peter Bruns and Heinz-Otto Luthe, 585–613.Wiesbaden: Harrassowitz Verlag, 2013.

Wheatley, Paul. The Places Where Men Pray Together. Cities in Islamic Lands, Seventh through the Tenth Centuries. Chicago: University of Chicago Press, 2001.

#VISUALIZATION

The Oxford English Dictionary defines #Visualization as the ability to form a mental picture as well as the actual process of creating something visible. As such, it covers the imagination as well as the production of a visual medium. Neither the reason nor the means of such an exercise are an issue. *

In the context of digital humanities, the 'translation' from one system of information or meaning into another to (re)present meaning seems the central point of #Visualization.

The definitions contributed by the mixed methods projects agree that "(t)he graphical, non-syntactic presentation of facts and relationships can help to facilitate a better understanding of complex information." They point out that there is a "reduction of information aspects [necessary] to achieve a meaningful clarity" (ArchiMediaL) and that the #Visualization is not a simple by-product of data collection but the result of various translation efforts and thus a consciously designed statement (ANCI). As such, the criteria and focus chosen for the visualizations have a critical impact on the way the information is communicated finally and they should foster the interaction concerning #Modelling as well. In addition, the interpretations derived from the #Visualization need inclusion into source critique.

This is not so crucial for projects that operate with visual media to start with, but for those for which the transfer of non-visual material into visual information is a major benefit of applying digital humanities: "Historical research with a visual dimension wins new perspectives in terms of new empirical insights and new analytical questions. Especially the compilation and visualization of data from very different sources enables new avenues of research not possible by the linear narrative or argumentation. Visualization of space and time enables the historian to see his/her data in a way he/she should not narrate them. It also provokes new research problems and new problems of representation as visualizations become facts of their own. More than the linguistic representation the visual representation may be read uncritically as given fact. It may also cause unwanted interpretations by the viewer which need to be taken into account" (DhiMu).

* "visualization, n.". in: *Oxford English Dictionary (OED)*, first published 1920; most recently modified version published online December 2021, https://www.oed.com/ [accessed: 20.05.2022].

Title: ArchiMediaL: Developing Post-Colonial Interpretations of Built Form through Heterogeneous Linked Digital Media
Team: Victor de Boer and Ronald Siebes (VU Amsterdam), Jan van Gemert, Carola Hein, Seyran Khademi and Tino Mager (TU Delft), Beate Löffler (Dortmund University of Technology), Dirk Schubert (HafenCity Universität Hamburg)
Corpus: Photographs, prints, drawing of architecture and built environment from Amsterdam city Archive. Google and Mapillary street view images.
Field of Study: Architectural history, Urban studies, Art history, Computer Vision
Institution: TU Delft, VU Amsterdam, Dortmund University of Technology, HafenCity Universität Hamburg
Technology: Artificial intelligence, Linked Data
Methods: Computer vision, deep learning, image recognition, similarity learning, cross-domain retrieval, crowdsourcing
Tools: ArchiMediaL Annotator

Computer Vision and Architectural History at Eye Level: Mixed Methods for Linking Research in the Humanities and in Information Technology (ArchiMediaL)

Tino Mager, Seyran Khademi, Ronald Siebes, Jan van Gemert, Victor de Boer, Beate Löffler, Carola Hein

Abstract *Information on the history of architecture is embedded in our daily surroundings, in vernacular and heritage buildings and in physical objects, photographs and plans. Historians study these tangible and intangible artefacts and the communities that built and used them. Thus valuable insights are gained into the past and the present as they also provide a foundation for designing the future. Given that our understanding of the past is limited by the inadequate availability of data, the article demonstrates that advanced computer tools can help gain more and well-linked data from the past. Computer vision can make a decisive contribution to the identification of image content in historical photographs. This application is particularly interesting for architectural history, where visual sources play an essential role in understanding the built environment of the past, yet lack of reliable metadata often hinders the use of materials. The automated recognition contributes to making a variety of image sources usable for research.*

Introduction

Architectural history is a discipline dedicated to the analysis of the built environment of the past. By examining historical buildings and sources of reference such as relics, texts and images, architectural history seeks, among other things, to uncover the conditions and characteristics of the structures of olden times. The buildings themselves, both surviving and lost, and the textual and visual sources that document their production, decay, and use are primary sources. Computer vision offers a potential to expand the range of classical approaches to extracting knowledge from

these sources.[1] While the automatic analysis of text and the search for key terms are well advanced, the automated recognition of image content is an area in which significant progress has only recently been made to such an extent that it appears applicable to visual archive material as well.[2]

Architectural production and also the documentation of the built environment have left a wealth of visual material since the mid-nineteenth century. In particular, photographs collected in public and private repositories have largely remained unexplored. Digitization is a necessary step to facilitate access to this enormous stock of visual material and digital cataloguing is essential to provide comprehensive and efficient information about the existence and accessibility of the material. Digital catalogues such as *Europeana*, the *German Digital Library* or the *Digital Public Library of America* provide access to millions of digitized documents; and architectural history repositories, such as the *Repository of Colonial Architecture and City Planning*[3], contain further collections of visual resources, including historical images of architecture and urban planning. Their digital availability is a decisive aid for research, as they can be viewed and analysed without much effort. Besides, they enable conclusions as to whether it is relevant to inspect the original.

Digitization and digital tools help architectural historians go beyond their traditional, usually limited visual material—archival documents, physical collections or books. As they dig into a big new set of imagery—electronic repositories, crowdsourcing or web-scale datasets—they need to refine their theories and methods. The handling of huge and unfamiliar datasets may throw up questions that go beyond the traditional hermeneutic reading of text and images. They must understand code as a cultural practice and learn to see qualitative data as the result of abstract 'technocratic' sorting that relies on established interpretation systems. Innovation in computer technology, both in crowdsourcing and in AI, creates opportunities and challenges for urban and architectural history, notably the recognition of visuals in vast archives. Crowdsourcing metadata for historical images is an issue closely related to those of communication, mediatization and urban future.

Metadata is required to navigate the visual contents of the repositories. They enable one to search, sort or filter a corpus that is non-verbal. Therefore, the names of

---

1 Beate Löffler, Carola Hein and Tino Mager. "Searching for Meiji-Tokyo: Heterogeneous Visual Media and the Turn to Global Urban History, Digitalization, and Deep Learning", *Global Urban History* (20.03.2018). https://globalurbanhistory.com/2018/03/20/searching-for-meiji-tokyo-heterogeneous-visual-media-and-the-turn-to-global-urban-history-digitalization-and-deep-learning/.

2 Alex Krizhevsky, Ilya Sutskever and Geoffrey E Hinton. "ImageNet Classification with Deep Convolutional Neural Networks", *Advances in Neural Information Processing Systems* 2 (01.12.2012): 1097–1105.

3 TU Delft, University Library: Colonial architecture & town planning, http://colonialarchitecture.eu.

objects, places and persons related to the image contents as well as keywords, etc., are essential annotations that help find specific visual sources in repositories and relate collection holdings all around the globe. This metadata needs to be assigned to the images systematically and correctly. Defective, wrong or missing annotations mean that the image material may not be found at all and is lost for research unless someone corrects them. Unfortunately, however, historical image collections in particular often have incomplete metadata.[4] Researchers or collection experts can hardly take on the task of manual annotation as due to the abundance of the material they could only handle a tiny part of the media concerned. Even larger teams would not be able to maintain a relevant amount of material. Computer vision can make a significant contribution here. It can help humanities scholars to manoeuvre through the wealth of digitally available visual material by recognising its content, in our case historical buildings. However, artificial intelligence (AI)-based computer vision systems need precise training to be able to recognize and identify specific image content. Therefore, it is necessary to build high-quality datasets that are used for training the algorithms and for evaluating their performance. At this intersection between information technology and humanities research, crowdsourcing can be used as a connecting method to bring together knowledge in the field of architectural history with specific data demands relevant for AI research.[5] The article outlines the application of crowdsourcing in the research project *ArchiMediaL*, which is dedicated to the identification of buildings in historical photographs by using computer vision.[6]

Use of computers for the analysis of historical urban images

The research project ArchiMediaL explores the possibilities of using current information technologies to open up architectural and urban image repositories for research. It develops strategies for the automatic recognition of historic image content through computer vision. Recent advances made in data-driven computer vision have improved the ability of visual intelligent systems to infer complex semantic

4 Beate Löffler and Tino Mager. "Minor politics, major consequences—Epistemic challenges of metadata and the contribution of image recognition", *Digital Culture & Society* 6, 2 (2021): 221–238.

5 Johan Oomen and Lora Aroyo. "Crowdsourcing in the cultural heritage domain: opportunities and challenges", *Proceedings of the 5th International Conference on Communities and Technologies* (2011): 138–149.

6 Seyran Khademi, Tino Mager, Ronald Siebes, Carola Hein, Beate Löffler, Jan van Gemert, Victor de Boer and Dirk Schubert, *Research project ArchiMediaL – Enriching and linking historical architectural and urban image collections* (TU Delft, VU Amsterdam, TU Dortmund, HafenCity University Hamburg). https://archimedial.eu.

relationships. The performance of modern computer vision requires a large number of images that have already been annotated. It takes millions of annotated images with hundreds of thousands of object classes to teach a computer 'to see the world'. Such images are abundant in the digital age, but training AI to see the past is more complex. To teach computers to recognize architecture in historic photographs, they must first be shown what the world looked like before the advent of digital cameras, otherwise they will not be able to recognize and detect objects and semantics of the past.

Again, to teach computers about the past, a large number of images are needed for training so that they can do the desired visual task, for example object recognition. The brain-inspired computer systems, referred to as convolutional neural networks (CNNs),[7] extract effective features in the form of tensor representation by seeing.[8] These trained models are later used for inference on visual data. Ideally, CNN models learn general rules during the training process, so that when they apply those rules to data they have not seen before, they can produce accurate classifications. Such inductive learning is consistent with human intelligence where the learned skills are used in real-world scenarios which might not have been fully covered in the training period. The more correlated the training and the test scenarios are, the more effective the learning is, for both the human and the machine intelligence. Accordingly, once a CNN model is trained on high-quality natural contemporary images for visual information retrieval, it cannot effectively perform the task on illustrations, low-quality and blurred images, drawings or paintings that are often found in archival records. In short, to develop intelligent tools that can handle the latter, we need to train the CNN models for archival image collections or else we end up creating incompetent tools.[9]

While Machine Learning based computer vision is finding its applications in the humanities domain,[10] the application of computer vision for architectural history and targeting historic photographs have not been pursued so far.[11] For facilitating the opening up of large image sets, research must start in a fairly well-documented

---

7 Krizhevsky, Sutskever, Hinton, "ImageNet".

8 Yoshua Bengio, Aaron C. Courville and Pascal Vincent. "Unsupervised Feature Learning and Deep Learning: A Review and New Perspectives", *CoRR*, abs/1206.5538 (24.06.2012): 1–30.

9 cf. Melvin Wevers and Thomas Smits, "The Visual Digital Turn: Using Neural Networks to Study Historical Images", *Digital Scholarship in the Humanities* 35, 1 (April 2020): 194–207, https://doi.org/10.1093/llc/fqy085.

10 Cf. Melvin Wevers and Thomas Smits. "The Visual Digital Turn: Using Neural Networks to Study Historical Images." *Digital Scholarship in the Humanities* 35, 1 (April 2020): 194–207, https://doi.org/10.1093/llc/fqy085.

11 A similarity can hitherto only be found in the Urban Panorama research project: North Carolina State University: Urban panorama. https://www.visualnarrative.ncsu.edu/projects/urban-panorama/.

area of architectural and urban history, since the performance and reliability of the algorithm can only be tested if the topic to which it is applied is known. To enable a high recognition rate, the objects that we study, in this case the built environment, should not have changed too much compared to the situation captured in the images, and the images must provide sufficient metadata to enable the results to be verified. To meet this requirement, a database of images of a well-studied location with a sufficient number of metadata is needed. We found it in the *Beeldbank* repository of the Amsterdam City Archives—the world's largest city archive. The Beeldbank contains several hundred thousand images taken in the streets of Amsterdam since the nineteenth century, among them many are images of facades, buildings and streets.[12]

In order to identify the buildings in the historic photographs with the help of computers, one can compare the image content with a large number of current photographs of Amsterdam that contain geo-information. Such a repository can be found in online mapping services such as *Mapillary*.[13] In Mapillary, a large part of Amsterdam's buildings is captured and provided with address data. If a building in a historical photo can be identified through similarity with a building in a Mapillary photo, the location of the building in the historical photo and thus the building itself are identified. The core of the project is, therefore, to use AI to automatically identify buildings in historical photographs that are geolocated in the Mapillary repository. The training of state-of-the-art AI models on available historical image data repositories can effectively help computers to become more intelligent and expert in the domain of historical data; in return, researchers and librarians can make use of these models to interact optimally with the visual archives.

This mutual interaction between computer scientists and humanity researchers breaks the classical pattern of computer science *serving* other fields without a real reciprocal conversation between the two parties. The two disciplines can effectively pursue research at eye-level, thereby producing new results in architectural history as a field of the humanities, and in computer science by making a step towards interpretable AI. However, training the AI models requires reliable data. In this case it means image pairs of the same building, once on a historical photo and once on a Mapillary photo. With a large number of such image pairs, the algorithm must be trained to be able to assign one of the many Mapillary images to a historical photo. The street address of the historical building can then be determined by the geolocation of the Mapillary photo. Having said this, these image pairs must be created by humans and they must be reliable. The involvement of many people is helpful for this purpose, as in this way a larger data set (>1000 image pairs) can be created relatively easily.

12 Beeldbank Stadsarchief Amsterdam. https://archief.amsterdam/beeldbank/.

13 Mapillary. http://mapillary.com, 2020 [accessed: 04.06.2020].

Crowdsourcing image pairs

High-quality data sets with images and reliable metadata are required to train and evaluate the AI systems. Crowdsourcing is a key factor in the creation of these. Virtually at the interface between social science research and information technology support, crowdsourcing serves to create the necessary data sets in a time- and cost-efficient way. One of the first Web2.0 crowdsourcing successes, where the value of content is created through the self-organised collaboration of an online community was launched on 9 March 2000 – *NuPedia*, the predecessor of *Wikipedia*.[14] More recently, various projects use crowdsourcing as a way to gather training data for Machine Learning methods. Especially large scale and complex learning approaches Deep Learning require large numbers of such instances to perform well. Crowdsourcing platforms such as Amazon *Mechanical Turk* and *CrowdFlower* enable researchers to collect training data efficiently. Applications of crowdsourcing for machine learning range from the medical domain (e.g. for recognizing tumors in images[15,16]), to Internet of Things[17] to Digital Humanities. Examples of the latter include geographical classification of location names in historical texts[18] or analysis of film colours.[19]

A first step involves the creation or compilation of training data: We needed annotations of historical photographs to match with geolocated photographs of the same building in their current form, so that computer could study the architectural

14 Larry Sanger. "The Early History of Nupedia and Wikipedia: A Memoir." *Open Sources 2.0: The Continuing Evolution*, ed. Chris DiBona, Mark Stone, and Danese Cooper (O'Reilly Media, Inc. 2005): 307–38.

15 Larry Sanger, "The Early History of Nupedia and Wikipedia: A Memoir", in: Chris DiBona, Mark Stone and Danese Cooper (eds.): *Open Sources 2.0: The Continuing Evolution*, ed. (O'Reilly Media, Inc. 2005): 307–38.

16 Shadi Albarqouni, Christoph Baur, Felix Achilles and Vasileios Belagiannis, "Aggnet: Deep Learning from Crowds for Mitosis Detection in Breast Cancer Histology Images", IEEE transactions on medical imaging 35.5 (2016): 1313–1321.

17 Qingchen Zhang, Laurence Tianruo Yang, Zhikui Chen, Peng Li and Fanyu Bu, "An Adaptive Dropout Deep Computation Model for Industrial IoT Big Data Learning with Crowdsourcing to Cloud Computing", IEEE Transactions on Industrial Informatics 15, 4 (April 2019): 2330–2337, doi: 10.1109/TII.2018.2791424.

18 Benjamin Adams and Grant McKenzie, "Crowdsourcing the Character of a Place: Character-Level Convolutional Networks for Multilingual Geographic Text Classification", Transactions in GIS 22.2 (2018): 394–408.

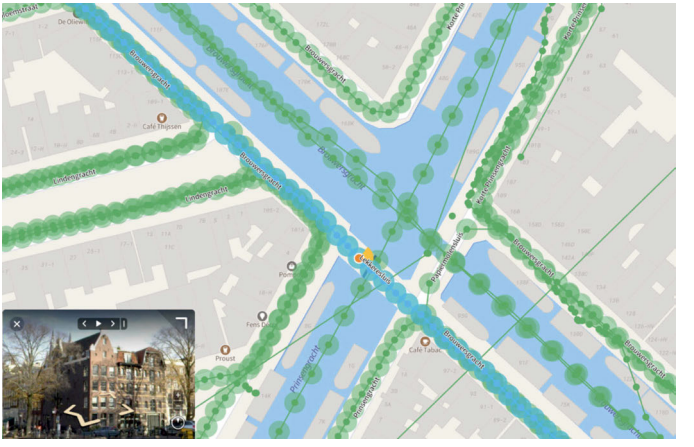
19 Barbara Flueckiger and Gaudenz Halter, "Building a Crowdsourcing Platform for the Analysis of Film Colors", Moving Image: The Journal of the Association of Moving Image Archivists 18.1 (2018): 80–83.

details. Some tasks are generic enough to be done via captchas,[20] but annotating historical street-view images requires some familiarity with the urban area depicted or some advanced training in planning-related fields. The same applies to the crowdsourcing task for the *ArchiMediaL* project. Due to limited means, it was necessary to find ways other than payment to motivate annotators to contribute. Therefore, the task needed to be uncomplicated and possibly combined with some gaming experience.[21]

The ArchiMediaL Annotator

In this regard, an annotation tool was created which showed the annotator a historical image and a 3D street-view navigator mostly positioned close to the expected location where the historical image had been taken, thus allowing the user to pan or zoom and match the historical and contemporary image in a game-like fashion (Figures 1 and 2).[22]

Fig. 1: Amsterdam, Brouwersgracht 160 on Mapillary. The green dots in the map indicate the positions of 360° images. Source: Mapillary (2016).

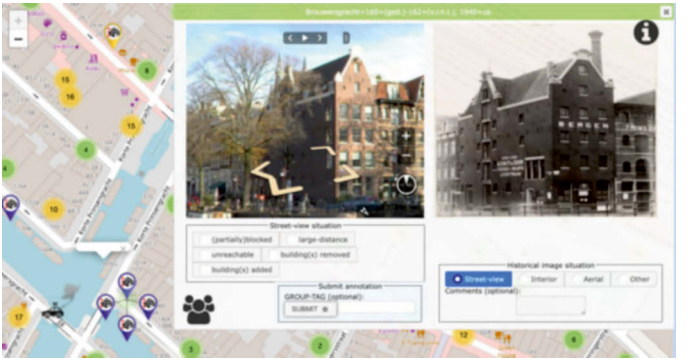


20 Luis von Ahn, Manuel Blum, Nicholas J. Hopper and John Langford, "Captcha: Using Hard AI Problems for Security", in: Eli Biham (ed.): *Advances in Cryptology—EUROCRYPT 2003* (Berlin, Heidelberg: Springer 2003), 294–311.

21 Benedikt Morschheuser, Juho Hamari and Jonna Koivisto, "Gamification in Crowdsourcing: A Review", 2016, 49th Hawaii International Conference on System Sciences (HICSS). IEEE, 2016.

22 Clark C, Serious games, 1987.

Fig. 2: A historical picture of the Brouwersgracht 160 in Amsterdam from the 1940s and the corresponding street view scene in the submission form of the crowdsourcing tool. Source: ArchiMediaL, 2020.



Once the participant navigates to the approximate location, (s)he can use the rotating, panning and zooming features to approximate the historical image. The participant can choose which area to discover and annotate by navigating on a map and then click on blue markers in the desired neighbourhood. Each marker contains one or more historical images that were taken in proximity of the marker. (An orange marker indicates that this historical image is already annotated and currently under review. If the marker is green, it means that at least one successful annotation is already added (see Figure 2). In case the historical image is not a street-view image (e.g. an interior or an aerial photo), the user can skip this task by selecting the appropriate checkbox and submit the result. Otherwise, the navigation procedure as described above can start and when finished, specific checkboxes can be marked which describe the current street-view situation in comparison to the situation in the historical photograph. Users can indicate whether a perfect match is not possible, for example, because new buildings have been built or old ones taken down or because the street-view panoramic tool cannot reach the point where the photographer of the historical image took the shot. The tool also provides an assessment interface that allows administrators to manually review or reject the results. This task takes only a fraction of the time allotted per image in comparison to the annotation task itself. The administrator can also make changes and resubmit a result. In the experimental setup, project administrators used this tool to verify or deny crowd submissions.

System design

To increase transparency and reusability of the tool, it must be Free and Open Source Software (FOSS). This requires a mix of technology, software libraries and data that in combination fulfils the requirements. For example, although they have high-quality content, Google Street-view data, professional web front-ends, and GIS platforms such as ArcGIS etc. have restrictive licences or are too expensive, thus rendering them unsuitable for the project. In the end, it was possible to put together a mix that enabled the implementation of the annotation platform. It contains the following elements that are briefly descried:

Mapillary is a street-level imaging platform that automatically annotates maps by using computer vision. Mapillary brings together a global network of contributors who want to make the world accessible to everyone by visualizing it and creating richer maps. In order to use their widgets, users must register for a free API key that can be accessed from their well-documented JavaScript library.[23] *OpenStreetMap* is an editable map database created and maintained by volunteers and distributed under the Open Data Commons Open Database License.[24] *Leaflet* is a widely used open source JavaScript library used to create web mapping applications.[25] Leaflet allows developers without a GIS background to easily view tiled web maps hosted on a public server with optional tiled overlays. It can load feature data from GeoJSON files, apply a style to them, and create interactive layers, such as markers with pop-ups when clicked. *Google Maps* provided the latitude and longitude coordinates that were used to map the Amsterdam Beeldbank title and description metadata on these coordinates. *PostGIS*[26] is a spatial database extender for PostgreSQL[27] object-relational database. It adds support for geographical objects and allows one to execute location queries in SQL. To store the locations internally in PostGIS, that is the latitude and longitude coordinates from the previous step, a transformation must be applied.[28]

Additionally, the following JavaScript libraries were used: jQuery is a JavaScript library designed to simplify HTML DOM tree traversal as well as manipulation, event handling, CSS animation, and *Ajax*.[29] Leaflet.*MarkerCluster* is used for clustering

23 The Mapillary API, 2020 [accessed: 04.06.2020].

24 https://www.openstreetmap.org, [accessed: 02.07.2021].

25 Vladimir Agafonkin et al., "Leaflet", https://leafletjs.com/, 2020 [accessed: June 2, 2020]; Leaflet Wikipedia. https://en.wikipedia.org/wiki/Leaflet (software), 2020 [accessed: May 28, 2020].

26 Postgis. https://postgis.net/, 2020 [accessed: 05.06.2020].

27 Postgresql. https://www.postgresql.org/, 2020 [accessed: June 5, 2020].

28 Postgresql-lat-lon.https://postgis.net/docs/STMakePoint. html, 2020 [accessed: June 5, 2020].

29 jquery. https://jquery.com/, 2020 [accessed: June 5, 2020].

markers in *Leaflet*.[30] It uses a grid-based clustering method which makes it ideal for providing a fast solution to the many markers problem. Tabulator creates interactive tables for any HTML Tables, JavaScript Arrays, *AJAX* data sources or JSON formatted data.[31] *FontAwesome* is an icon toolkit featuring icon font ligatures, an SVG framework, official NPM packages for popular front-end libraries such as React, and facilitating access to a new CDN.[32] Most of the *ArchiMediaL* annotator icons have *FontAwsome* as the source. Node.js is a platform built on Chrome's JavaScript runtime for easily building fast and scalable network applications.[33] *Node.js* serves as the larger part of the *ArchiMediaL* back-end. *Puppeteer* is a *Node.js* library which provides a high-level API to control Chrome or Chromium over the *DevTools* Protocol.[34] *Puppeteer* runs headless by default, but can be configured to run full (non-headless) Chrome or Chromium.

Data collection

Users

We have experimented with various strategies to attract new commentators to the platform. The easiest and most successful one is that which gives students some time to experiment during lectures. The peak in new subscriptions was reached during one of the lectures at TU-Delft in the week beginning December 16, 2019 (see Figure 3).

30 Leaflet.markercluster. https://github.com/Leaflet/Leaflet.markercluster, 2020 [accessed: June 5, 2020].

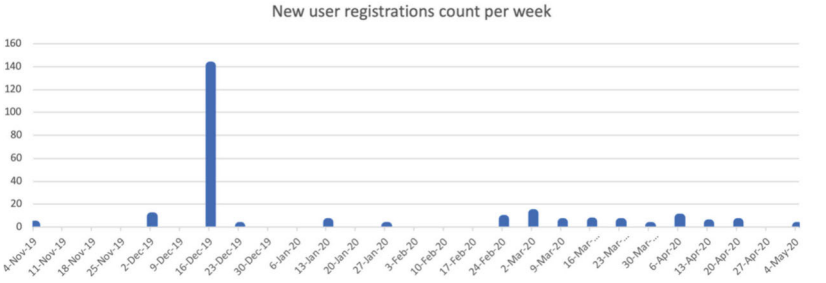
31 Tabulator. http://tabulator.info/, 2020 (accessed June 5, 2020).

32 Fontawesome. https://github.com/FortAwesome/Font-Awesome, 2020 (accessed June 5, 2020).

33 Node.js. https://nodejs.org/en/, 2020 (accessed June 5, 2020).

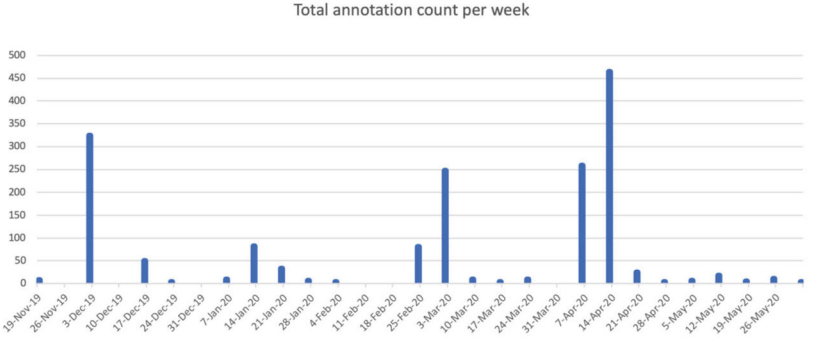
34 Puppeteer for node.js. https://github.com/puppeteer/puppeteer, 2020 (accessed June 5, 2020).

Fig. 3: New registrations per week.



Unfortunately, this stress test resulted in an unforeseen, unacceptable peak in the use of computing resources on the shared servers; and the host decided to temporarily shut down all processes. Since then, a considerable amount of time has actually been spent on making the processes more efficient. We hoped that the students would continue to play around with the annotator and help us with further annotations, but this did not happen. This explains why, although most new registrations were made during this event, most of the students did not submit an annotation either then or afterwards. Figure 4 shows the weekly number of new subscriptions on our platform. In the future, we plan to check the effectiveness of the measures we have taken to attract new subscribers.

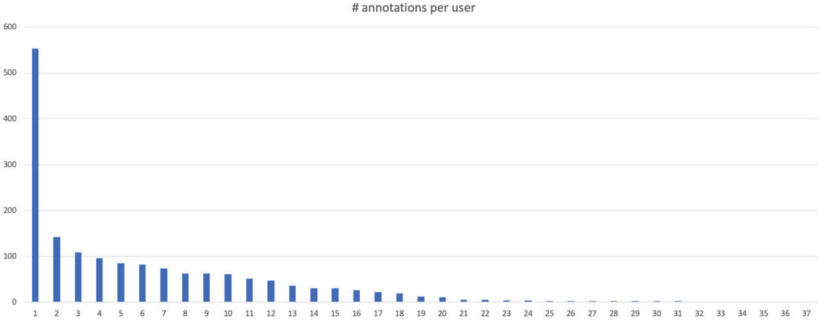
Fig. 4: Annotations per week



A highly distorted distribution, another interesting observation, can be found in Figure 5. Only a few users are responsible for the majority of the annotations. This is

a well-known phenomenon known as crowdsourcing participation inequality[35] best known in the case of contributions on Wikipedia.[36]

Fig. 5: Total amount of annotations per user.



Annotation statistics

Table 1 shows statistics of the annotations gathered during the initial experiment period (November 2019-May 2020). A total of 1,656 annotations were made, of which the majority (1,116 or 67%) were annotations where participants identified a historical image as a street view, with images of interiors being the second largest category (with 225 annotations). The 'other' category was the third largest. It included images, such as portraits and photographs of building equipment, paintings and other objects. The smallest category was that of aerial photographs (95).

Annotations submitted initially were categorized as 'pending review'. Thus, administrators could review and then accept or reject the annotations. In our experiment, the role of administrators was taken up by the authors. The table also shows the numbers of accepted versus rejected annotations. In total, 80% of annotations were accepted. This shows the quality of individual annotations. The percentage of accepted annotations was the lowest for street-view (75%) and highest for interiors (95%). One explanation for this is that street-view annotations are more elaborate as participants are asked to actually navigate to the correct location leading to greater risk of errors, whereas for the other categories a simple categorization is sufficient.

35 Stewart Osamuyimen, David Lubensky and Juan M. Huerta, "Crowdsourcing Participation Inequality: A SCOUT Model for the Enterprise Domain", Proceedings of the ACM SIGKDD Workshop on Human Computation. 2010.

36 Wikipedia. List of wikipedians by number of edits. https://en.wikipedia.org/wiki/Wikipedia:ListofW ikipediansbynumberofedits, 2020 (accessed: June 5, 2020).

Table 2 shows the result of further annotation details about the accepted street-view annotations by showing the additional categories identified by the participants. Here, we see that a small number of annotations concern street-view situations where the view of the building is partially hindered or changed. For the majority, this is however not the case (67.7%) and the feasibility of this annotation task is shown in 2/3rds of the cases in Amsterdam.

Table 1: Total of accepted, pending and rejected annotations per category.

	Street-view	Interior	Aerial	Other	Total
Accepted	835	243	84	165	1327
Rejected	278	12	11	25	326
Pending review	3	0	0	0	3
Total	1116	255	95	190	1656

Table 2: Statistics on street-view situations for accepted annotations.

Street-view situation	Total (percentage)
Blocked (partially)	103 (12.3%)
Large distance	50 (6.0%)
Unreachable	45 (5.4%)
Buildings removed	60 (7.2%)
Buildings added	79 (9.5%)
None	565 (67.7%)

Similarity Learning

Recent research literature shows that similarity learning is a powerful way to gain insight into data.[37] Considering the large variety of objects and therefore classes in most archival data and, in our case, the cross-time dataset for historical and current street view of Amsterdam, we use deep similarity learning for representation learning. We address the cross-domain image retrieval task, formulated as content-based image retrieval (CBIR), where the semantic similarity needs to be learned to find the most similar images in the gallery with reference to the query image. We use a *Siamese network* that uses two sister networks with shared learning parameters for the training process. The training tuples are images and their labels, in contrast to the classification where the training pair is the image and the label. The CNN network learns to project an image to a vector (latent) space such that similar images are placed closer, in terms of Euclidean distance, compared to dissimilar images.[38]

Once the network is trained for similarity learning, the CNN is used to map all the images in the dataset to a vector space. The distances between the vector representations in Euclidean sense determine the corresponding image pairs. In the retrieval task, the distance between the vector representations of the query image is computed with regard to the vector representations of the images in the gallery. The images are then ranked in ascending order with respect to their distances to the query.

Cross-domain Retrieval

In the context of cross-domain CBIR, representations of the objects in the gallery database can be potentially different from the query image. For example, the images may contain sketches, paintings or old photos as discussed in the beginning.

---

37 Sumit Chopra, Raia Hadsell and Yann Lecun, "Learning a similarity metric discriminatively, with application to face verification", *The IEEE Computer Society Conference on Computer Vision and Pattern Recognition CVPR 2005*, 1 (07, 2005): 539– 546; Gregory Koch, Richard Zemel and Ruslan Salakhutdinov, "Siamese Neural Networks for One-Shot Image Recognition", *ICML deep learning workshop* 2. 2015.

38 In other words, deep CNN is a mapping function f from an image of size w × h with r color channels to a k-dimensional representation space, $f : IN^{w\times h\times r} IR^k$ where distances in IR^k between similar image pairs are smaller than distances to dissimilar image pairs by a predefined margin m in the desired metric space, i.e., dPositive + m ≤ dNegative, where dPositive = d(f(I),f(I')|y = 1), (1) dNegative = d(f(I), f(I')|y = 0). (2) Similar image pairs have a label y = 1 and dissimilar image pairs the label y = 0. We consider a Euclidean metric space where d(f(I),f(I')) outputs the Euclidean distance between two vectors of representations for an image pair {I, I'}.

For cross-domain CBIR tasks, a common failure mode, which comes into play while deploying the CNN models trained only to detect single-domain images, is that the network places similar-style images in a neighbourhood which depicts different objects as it has never seen one from the second domain (featuring historical images here). To resolve this domain-disparity issue, we propose to learn domain-invariant image representations that focus on semantics rather than on image style or colour. This leads to a specialized CNN model that learns indifferent image representations for archival and contemporary image domains but is discriminative at the content level.

In our work, a combination of attention and domain adaption is used to train robust CNN networks for an age-agnostic image retrieval task.[39] Here, similarity is learned by training a Siamese network to detect images with the same geo-tags in the contemporary image domain. It is commonly referred to as weakly supervised learning as the labels for training are not the same as the ones for testing. In our case, the test (evaluation) set contains a query from historical Amsterdam, but the gallery houses current street-views of Amsterdam. The results reveal a performance gap as the intra-domain trained model for street-view images is tested on cross-domain data, indicating a drop from 99% top 20 accuracy to 28%. The main conclusion is that full supervision is required to achieve reasonable performance for the cross-domain image retrieval task.

Conclusion

The research project started out by exploring automatic recognition of buildings in historic images by AI. Analogous to automatic facial recognition, buildings are to be recognized and identified. The input objects are historical images of buildings whose contents are localized by a specially designed and trained artificial neural network. The localization allows unique identification. The recognition can be realized for a specific area by providing the computer with many images of already identified, that is, localized buildings. This training of the network enables the computer to recognize buildings in historical images hitherto unknown to it. First, however, the training dataset must be created with hundreds of historical images of identified buildings.

Computer vision can help to identify image content in historical images of the built environment. As the case of Amsterdam exemplifies, a stock of +400k architectural images from the Beeldbank archives has not yet been clearly identified in

39 Ziqi Wang, Jiahui Li, Seyran Khademi and Jan van Gemert. "Attention-aware Age-agnosticVisual Place Recognition", *The IEEE International Conference on Computer Vision (ICCV) Workshops* (October 2019).

terms of geographical location, building name or type. Millions of similar images exist worldwide in online and offline repositories. The identification of the image content is not possible for architectural and urban historians simply because of the sheer quantity of the images. As a result, valuable visual source material for research in the humanities is lost. Computer vision seems to be helpful here. As buildings can be identified in images by comparing each of them with a geolocated image of the same building, street views from Mapillary are used as a reference system. However, there is no one-size-fits-all solution when it comes to learning representative features for various visual domains.[40] A CNN model pre-trained to follow standard benchmarks will not work on a non-standard image dataset with a different style of images and sizes.[41]

In our experimental setup mentioned, we showed that unsupervised or weakly supervised methods performed poorly when it came to recognizing historical image sets. Therefore, some form of supervised learning is inevitable for feature learning. The contribution of supervised learning to the cross-domain image retrieval is only revealed once annotated data is available for training. This type of crowdsourcing is a valuable way to achieve a sufficient number of exactly matching image pairs. The crowdsourcing stands at the intersection of deep learning and humanities research. Through advertising in university courses and social media, volunteers helped pair 1,656 images. A review and correction of these pairs have resulted in a set of 902 useful image pairs that can be used to train the neural network that is currently in progress. We hope to achieve a useful level of accuracy which will make automatic image content recognition a useful tool for architectural history research.

The findings of the ArchiMediaL project open up new perspectives for architectural history in diverse areas. Researchers, politicians and planners can explore 4D reconstructions of the past (e.g., in their respective websites, the HistStadt4D research project or various local Time Machine projects) to increase historical understanding, to enrich tourist experiences, or to facilitate design decisions. New research questions can be framed through the availability of such data. Using the ArchiMediaL tool can raise numerous questions. For example, scholars could examine bubbles where data are more or less available for raising questions such as: How

40 Xiaohua Zhai, Joan Puigcerver, Alexander Kolesnikov, Pierre Ruyssen, Carlos Riquelme, Mario Lucic, Josip Djolonga, Andre Susano Pinto, Maxim Neumann, Alexey Dosovitskiy, Lucas Beyer, Olivier Bachem, Michael Tschannen, Marcin Michalski, Olivier Bousquet, Sylvain Gelly and Neil Houlsby, "A large-scale study of representation learning with the visual task adaptation benchmark", 2019.arXiv:1910.04867.

41 Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A. Wichmann and Wieland Brendel, "Imagenet-trained CNNs [PLS CHECK THE CAPITALIZATION] are Biased towards Texture; Increasing Shape Bias Improves Accuracy and Robustness", *CoRR*, *abs/1811.12231*, 2018.

does the availability of pictorial data from the past correlate with the architectural quality of the building stock or the socioeconomic composition of its citizens?

In the case of Amsterdam, many datasets with spatial information are available in digital form, including the ones based on the age of buildings, the number of breeding birds in green areas, climate information (heat, drought, flooding), post-war monumental wall art and land value, to name but a few. The crossing of this data with the visual sources localized within the work of the project allows framing of new research questions that investigate the connection between architectural or urban form and phenomena such as property value or gentrification. An expansion to other cities and areas will make it possible to formulate new findings on the basis of a greater number of correlations and thus make more general statements than those emerging from individual case studies.

The application of AI in historical research is not a mere information technology task. As in any mixed-methods approach, it requires meeting and communicating with different disciplines, and profound expertise in the humanities.[42] Interpretation of the past needs careful framing of available data to achieve meaningful findings. Such a step can only be made through interdisciplinary collaboration among humanities scholars, computer scientists, historians and designers. Moreover, this project has required people to contribute their knowledge both to create the training dataset and to eventually evaluate the performance of the algorithm. Crowdsourcing can offer an important opportunity for participation—important in terms of not only identifying past worlds, but also involving people in research. It would help integrate their points of view and ultimately awaken their interest in questions of urban history and urban development.

In Digital Humanities, the heterogeneity of data demand and data supply is a common challenge. Since research in the field of information technology requires high quality datasets, for example, to advance neural networks and the performance of computer vision, it is common to use datasets that are well-suited for technological search and to ignore those that fall short of the task. These are the datasets that humanities work with. In the course of addressing this challenge, we developed tools to enhance, even create the required datasets. In doing so, we were able to identify research questions that were useful for both scientific parties. Thus we refused to make IT research a mere supplier of a solution to a humanities problem and instead launched a joint research initiative involving researchers from all disciplines. In such an intellectual environment, the combination of humanities and IT research can help to make progress and gain new insights in both areas. It also offers the opportunity to formulate new research questions and to advance to more complex interdisciplinary research designs.

42 John W. Creswell and Vicki L. Plano Clark, Designing *and Conducting Mixed Methods Research*. 3rd. ed. (Los Angeles: Sage, 2018).

Bibliography

Abt, Clark C. Serious games. Lanham et al.: University Press of America, 1987.

Adams, Benjamin, and Grant McKenzie. "Crowdsourcing the Character of a Place: Character-Level Convolutional Networks for Multilingual Geographic Text Classification". In Transactions in GIS 22.2 (2018): 394–408.

Bengio, Yoshua; Aaron C. Courville, and Pascal Vincent. "Unsupervised Feature Learning and Deep Learning: A Review and New Perspectives". In CoRR, abs/1206.5538 (24.06.2012): 1–30.

Chopra, Sumit; Raia Hadsell, and Yann Lecun. "Learning a similarity metric discriminatively, with application to face verification". In The IEEE Computer Society Conference on Computer Vision and Pattern Recognition CVPR 2005, 1 (07, 2005): 539–546.

Creswell, John W., and Vicki L. Plano Clark. Designing and Conducting Mixed Methods Research. 3rd. ed. Los Angeles: Sage, 2018.

Flueckiger, Barbara; Gaudenz Halter. "Building a Crowdsourcing Platform for the Analysis of Film Colors". In Moving Image: The Journal of the Association of Moving Image Archivists 18.1 (2018): 80–83.

Geirhos, Robert; Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A. Wichmann, and Wieland Brendel. "Imagenet-trained CNNs are Biased towards Texture; Increasing Shape Bias Improves Accuracy and Robustness". In CoRR, abs/1811.12231, 2018, 22 p.

Khademi, Seyran, Tino Mager, Roland Siebes, Carola Hein, Beate Löffler, Jan van Gemert, Victor de Boer, and Dirk Schubert. Research project ArchiMediaL – Enriching and linking historical architectural and urban image collections (TU Delft, VU Amsterdam, TU Dortmund, HafenCity University Hamburg). https://archimedial.eu.

Koch, Gregory, Richard Zemel, and Ruslan Salakhutdinov. "Siamese Neural Networks for One-Shot Image Recognition". In ICML deep learning workshop 2. 2015.

Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E. Hinton. "ImageNet Classification with Deep Convolutional Neural Networks". In Advances in Neural Information Processing Systems 2 (01.12.2012): 1097–1105.

Löffler, Beate, Carola Hein, Tino Mager. "Searching for Meiji-Tokyo: Heterogeneous Visual Media and the Turn to Global Urban History, Digitalization, and Deep Learning". In Global Urban History (20.03.2018). https://globalurbanhistory.com/2018/03/20/searching-for-meiji-tokyo-heterogeneous-visual-media-and-the-turn-to-global-urban-history-digitalization-and-deep-learning/.

Löffler, Beate, and Tino Mager. 'Minor politics, major consequences—Epistemic challenges of metadata and the contribution of image recognition'. In Digital Culture & Society 6, 2 (2021): 221–238.

Mager, Tino, and Carola Hein. "Mathematics and/as Humanities: Linking Humanistic Historical to Quantitative Approaches". In The Mathematics of Urban Morphology, edited by Luca D'Acci. 523–528. Cham: Birkhäuser, 2019.

Morschheuser, Benedikt, Juho Hamari, Jonna Koivisto. "Gamification in Crowdsourcing: A Review". In Proceedings of the 49th Hawaii International Conference on System Sciences (HICSS). IEEE, 2016.

Oomen, Johan, and Lora Aroyo. "Crowdsourcing in the cultural heritage domain: opportunities and challenges". In Proceedings of the 5th International Conference on Communities and Technologies (2011): 138–149.

Sanger, Larry. "The Early History of Nupedia and Wikipedia: A Memoir". In Open Sources 2.0: The Continuing Evolution, edited by Chris DiBona, Mark Stone, and Danese Cooper, 307–338. Beijing et al.: O'Reilly Media, Inc., 2005.

Albarqouni, Shadi, Christoph Baur, Felix Achilles, and Vasileios Belagiannis. "Aggnet: Deep Learning from Crowds for Mitosis Detection in Breast Cancer Histology Images". In IEEE transactions on medical imaging 35.5 (2016): 1313–1321.

Stewart, Osamuyimen, David Lubensky, and Juan M. Huerta. "Crowdsourcing Participation Inequality: A SCOUT Model for the Enterprise Domain". In Proceedings of the ACM SIGKDD Workshop on Human Computation. 2010.

von Ahn, Luis, Manuel Blum, Nicholas J. Hopper, and John Langford. "Captcha: Using Hard AI Problems for Security." In Advances in Cryptology—EUROCRYPT 2003, edited by Eli Biham, 294–311. Berlin, Heidelberg: Springer 2003.

Wang, Ziqi; Jiahui Li, Seyran Khademi, and Jan van Gemert. "Attention-aware Age-agnostic Visual Place Recognition." In The IEEE International Conference on Computer Vision (ICCV) Workshops. (Oct 2019).

Wevers, Melvin, and Thomas Smits. "The Visual Digital Turn: Using Neural Networks to Study Historical Images". In Digital Scholarship in the Humanities 35, 1 (April 2020): 194–207, https://doi.org/10.1093/llc/fqy085.

Wikipedia. List of wikipedians by number of edits. https://en.wikipedia.org/wiki/Wikipedia:ListofW ikipediansbynumberofedits, 2020 [accessed: 05.06.2020].

Zhai, Xiaohua; Joan Puigcerver, Alexander Kolesnikov, Pierre Ruyssen, Carlos Riquelme, Mario Lucic, Josip Djolonga, Andre Susano Pinto, Maxim Neumann, Alexey Dosovitskiy, Lucas Beyer, Olivier Bachem, Michael Tschannen, Marcin Michalski, Olivier Bousquet, Sylvain Gelly, and Neil Houlsby. "A large-scale study of representation learning with the visual task adaptation benchmark". 2019. arXiv:1910.04867

Zhang, Qingchen; Laurence Tianruo Yang, Zhikui Chen, Peng Li, and Fanyu Bu. "An Adaptive Dropout Deep Computation Model for Industrial IoT Big Data Learning With Crowdsourcing to Cloud Computing". In IEEE Transactions on Industrial Informatics 15, 4 (April 2019): 2330–2337, doi: 10.1109/TII.2018.2791424.

Repositories

TU Delft, University Library: Colonial architecture & town planning, http://colonialarchitecture.eu.
North Carolina State University: Urban panorama. https://www.visualnarrative.ncsu.edu/pro jects/urban-panorama/.
Beeldbank Stadsarchief Amsterdam. https://archief.amsterdam/beeldbank/.
Mapillary. http://mapillary.com, 2020 [accessed: 04.06.2020]

#CANON

The OED defines #Canon (in the sense of hierarchies or cultural norms), among others, as a "standard of judgement or authority; a test, criterion, [or] means of discrimination" and as an extension of its original ecclesiastical meaning, that is, "writings of a secular author accepted as authentic." *

As such, the #Canon is a fundamentally qualitative phenomenon that refers to intra-cultural characteristics forged by history and society. Its discussion mirrors the awareness about the role of epistemic systems and research environments within academic disciplines and fields of research especially within the humanities (see #Corpus). Here, "controversies about issues of the canon abound" and while some developments in digital humanities celebrate the icons and their legacy, many digital humanities projects such as those of mixing methods "aim to transcend the canon of the usual suspects (the so-called 'great masters')" (BachBeatles) thus moving towards a comprehensive understanding of the phenomena at hand and the broadening of the #Canon. In addition, since "canons of knowledge form a reference framework against which research ideas and results are evaluated together with worldviews," they are "a backbone of any research, a means to ensure quality standards and a possible obstruction as soon as new ideas step 'too' far out" (ArchiMediaL). As such, canons are a means to articulate conceptual belonging. In the context of digital humanities, especially in the process of mixing methods the negotiation of contrasting canons carries "both the possibility of overcoming outdated patterns and the risk of losing structure to meaninglessness" (ArchiMediaL).

* "Canon, n.1.". in: *Oxford English Dictionary (OED)*, first published 1888; most recently modified version published online March 2022 with a draft addition from 2002, https://www.oed.com/ [accessed: 20.05.2022].

Title: From Bach to the Beatles: Exploring Compositional Building Blocks and Musical Style Change with Hermeneutic and Computational Methods
Team: Martin Rohrmeier, Markus Neuwirth, Christoph Finkensiep, Ken Déguernel
Corpus: Instrumental Music between ca. 1700 and 1970
Field of Study: Computational Musicology
Institution: Digital and Cognitive Musicology Laboratory, École Polytechnique Fédérale de Lausanne
Methods: Various corpus research methods (adapted from NLP), pattern recognition, skipgrams, computational modelling
Tools: Schema-annotation app
Technology: Julia, ClojureScript, Verovio

Musical Schemata: Modelling Challenges and Pattern Finding (BachBeatles)

Markus Neuwirth, Christoph Finkensiep, and Martin Rohrmeier

Abstract *'From Bach to the Beatles: Exploring Compositional Building Blocks and Musical Style Change with Hermeneutic and Computational Methods' aims at finding voice-leading schemata in digital corpora, analyzing their characteristics, and describing their chronological distributions between ca. 1700 and 1970. Voice-leading schemata are fundamental building blocks of Western music across historical periods and styles, ranging from Renaissance, Baroque, and Classical to modern tonal music. Rather than focusing on concrete findings, the chapter adopts a meta-perspective, explaining and discussing the methodological choices regarding initial hypotheses, modelling options, encodings, pattern-matching methods, and the interpretation frameworks. The final part of the chapter is devoted to a discussion of how computational methods and hermeneutics can interact with one another in a productive manner.*

Computational music theory

Music theory is primarily concerned with characterizing musical structures as they occur across cultures, historical periods, and styles, aiming to reveal the general principles underlying these structures. Previous music-theoretical scholarship has suffered to some extent from a number of core problems including the problem of intuitive statistics, a lack of methodological transparency, and various sampling biases.[1] These shortcomings have been addressed in the more recent field of musical corpus studies. Over the past two decades, music theory has intensified efforts to adopt powerful digital/computational methods in order to empirically scrutinize armchair hypotheses about the nature of musical structures and their historical de-

1 See, for instance, Markus Neuwirth and Martin Rohrmeier, "Wie wissenschaftlich muss Musiktheorie sein? Chancen und Herausforderungen musikalischer Korpusforschung", *Zeitschrift der Gesellschaft für Musiktheorie* 13/2 (2016), 171–193.

velopment, relying on a steadily growing amount of datasets.[2] These datasets can be human- or machine-generated, refer to structural domains such as harmony, counterpoint, melody, rhythm, meter, timbre etc., and involve Western as well as non-Western musical styles.[3]

However, it is only a more recent trend in digital music research to consider more complex musical objects such as voice-leading schemata[4], cadences[5], or sonata form[6]. Successfully dealing with these phenomena not only crucially depends on expert knowledge, but also presents the challenge of how to model and operationalize high-level music-theoretical concepts in ways suitable for digital

2 Several societies and interest groups have been launched that testify to this trend, most importantly the Music Encoding Initiative (MEI) in 2005 and the International Society for Music Information Retrieval (ISMIR) in 2008.

3 E.g., Christof Weiß, Frank Zalkow, Vlora Arifi-Müller, Meinard Müller, Hendrik Vincent Koops, Anja Volk and Harald G. Grohganz, "Schubert Winterreise Dataset: A Multimodal Scenario for Music Analysis", *ACM Journal on Computing and Cultural Heritage* (JOCCH), 14/2 (2021), 1–18.

4 See, for instance, James Symons, "A Cognitively Inspired Method for the Statistical Analysis of Eighteenth-Century Music, as Applied in Two Corpus Studies", PhD thesis, Northwestern University, 2017; David R. W. Sears, Marcus T. Pearce, William E. Caplin and Stephen McAdams, "Simulating Melodic and Harmonic Expectations for Tonal Cadences Using Probabilistic Models", *Journal of New Music Research* 47/1 (2018), 29–52; David R. Sears and Gerhard Widmer, "Beneath (or Beyond) the Surface: Discovering Voice-leading Patterns with Skip-grams", *Journal of Mathematics and Music* 15/3 (2020), 1–26; DOI: 10.1080/17459737.2020.1785568; Christoph Finkensiep, Markus Neuwirth and Martin Rohrmeier (2018), "Generalized Skipgrams for Pattern Discovery in Polyphonic Streams", in: E. Benetos, E. Gómez, X. Hu, and E. Humphrey (eds.): *Proceedings of the 19th International Society for Music Information Retrieval Conference (ISMIR)* (Paris 2018) 547–553; Andreas Katsiavalos, Tom Collins and Bret Battey, "An Initial Computational Model for Musical Schemata Theory", in: *Proceedings of the International Society on Music Information Retrieval*, 2019, 166–172; and Christoph Finkensiep, Ken Déguernel, Markus Neuwirth and Martin Rohrmeier, "Voice-Leading Schema Recognition Using Rhythm and Pitch Features", in: *Proceedings of the International Society for Music Information Retrieval* (ISMIR) (Montreal 2020).

5 For instance, Ben Duane, "Melodic Patterns and Tonal Cadences: Bayesian Learning of Cadential Categories from Contrapuntal Information", *Journal of New Music Research* 48/3 (2019), 197–216; and Johannes Hentschel, Markus Neuwirth and Martin Rohrmeier, "The Annotated Mozart Sonatas: Score, Harmony, and Cadence", *Transactions of the International Society for Music Information Retrieval* 4/1 (2021), 67–80; https://doi.org/10.5334/tismir.63

6 Christof Weiß, Stephanie Klauk, Mark Gotham, Meinard Müller and Rainer Kleinertz, "Discourse Not Dualism: An Interdisciplinary Dialogue on Sonata Form in Beethoven's Early Piano Sonatas", in: *Proceedings of the International Society for Music Information Retrieval Conference* (ISMIR) 2020, 199–206; Pierre Allegraud, Louis Bigo, Laurent Feisthauer, Mathieu Giraud, Richard Groult, Emmanuel Leguy and Florence Levé, "Learning Sonata Form Structure on Mozart's String Quartets", *Transactions of the International Society for Music Information Retrieval* 2/1 (2019), 82–96.

and empirical research. Our project titled 'From Bach to the Beatles: Exploring Compositional Building Blocks and Musical Style Change with Hermeneutic and Computational Methods' aimed (1) to address that modelling challenge and (2) to come up with algorithmic solutions to finding voice-leading schemata in digital corpora.

Why study musical schemata?

Why is the study of voice-leading schemata in particular a worthwhile endeavour? Firstly, most voice-leading schemata are comparatively easy to recognize by both trained and untrained listeners, and they play a pivotal educational role in ear-training classes at music universities. Secondly, voice-leading schemata can be seen as an important musical object from the point of view of diachronic style development. They are frequently used patterns to be found across historical periods, ranging from Renaissance, Baroque, and Classical to modern tonal music. Examples include such well-known and cognitively salient schemata as the Lamento, the Pachelbel, the descending-fifths sequence, and various forms of cadences (i.e., closing formulae). Some schemata even carry with them specific expressive or affective connotations: the Lamento, for instance, is traditionally associated with grief and depression; similarly, the Teufelsmühle (or Omnibus) conveys notions of uncertainty, horror, and imminent death.[7]

Finally, voice-leading schemata constitute a rich musical object in view of both their structure and usage. A schema serves as an abstract contrapuntal template that composers can elaborate in multiple ways by flexibly inserting embellishing notes. Apart from the frequency of use of particular schemata, it is the concrete way in which schemata are instantiated in a composition that helps distinguish musical styles from one another.

Our 'From Bach to the Beatles' project was primarily concerned with finding voice-leading schemata in large digital corpora in order to be able to answer questions about the use of schemata over time and geographical areas. At present, there is only scant quantitative evidence of the frequency and diachronic distribution of

7 E.g., William E. Caplin, "Topics and Formal Functions: The Case of the Lament", in: Danuta Mirka (ed.): *The Oxford Handbook of Topic Theory* (New York: Oxford University Press, 2014), 415–452; John A. Rice, "The Morte: A Galant Voice-Leading Schema as Emblem of Lament and Compositional Building-Block", *Eighteenth-Century Music* 12/2 (2015), 157–181; Paula J. Telesco, "Enharmonicism and the Omnibus Progression in Classical-Era Music", *Music Theory Spectrum* 20/2 (1998), 242–279; and Marie-Agnes Dittrich, '"Teufelsmühle" und "Omnibus"', *Zeitschrift der Gesellschaft für Musiktheorie* 4/1–2 (2007), 107–121, https://doi.org/10.31751/247

schemata across history;[8] large-scale, machine-readable datasets on schemata are almost non-existent.[9] For assessing the prevalence of schemata in musical corpora, automated recognition of schema instances can be a time- and cost-efficient alternative to manually labelled data.

There are, however, two key challenges that computational approaches face as they seek to detect note patterns in music: (1) the multidimensional (polyphonic) structure of music as opposed to, for example, the sequential structure of written text; and (2) the highly flexible nature of these patterns, given that the structural notes in the individual voices can be elaborated in a great variety of ways. These problems will be explained in detail further below (see section The challenge of finding schemata).

What are musical schemata?

The schema concept itself comes from cognitive psychology and has subsequently been introduced in the field of music research.[10] Related concepts used in this context are script, frame, prototype, idealtype, archetype, model, meme, idiom (in the sense defined by construction grammar), cognitive map, and associative network.[11] In principle, any music-structural domain (e.g., harmony, form, and counterpoint) can be conceptualized in terms of schemata. Nonetheless, the term 'schema' has cur-

8 Among the exceptions are Robert O. Gjerdingen, *A Classic Turn of Phrase: Music and the Psychology of Convention* (Philadelphia: University of Pennsylvania Press, 1988); Gjerdingen, *Music in the Galant Style* (New York: Oxford University Press, 2007); Vasili Byros, "Towards an 'Archaeology' of Hearing: Schemata and Eighteenth-century Consciousness", *Musica Humana* 1/2 (2009), 235–306; and Byros, "Trazom's Wit: Communicative Strategies in a 'Popular' Yet 'Difficult' Sonata", *Eighteenth-Century Music* 10/2 (2013), 213–252.

9 For a recent contribution to this issue, see Finkensiep et al., "Voice-Leading Schema Recognition".

10 The cognitive notion of schema is described in David E. Rumelhart, "Schemata: The Building Blocks of Cognition", in: Rand J. Spiro, Bertram C. Bruce, and William F. Brewer (eds.): *Theoretical Issues in Reading Comprehension* (Hillsdale, NJ: Lawrence Erlbaum, 1980), 33–58. Arguably the first author to explicitly apply the concept of schema to musical phenomena was Leonard B. Meyer (e.g., Meyer, "Innovation, Choice, and the History of Music", *Critical Inquiry* 9/3 [1983], 517–544); it was picked up by Robert Gjerdingen in 1988.

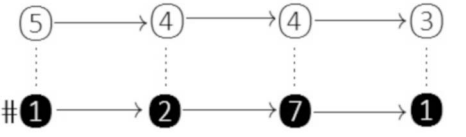
11 The notion of 'script' is introduced in Roger Schank and Robert P. Abelson, *Scripts, Plans, Goals, and Understanding: An Inquiry into Human Knowledge Structures* (New York: Halsted, 1977). For a Neo-Darwinian, memetic approach to schemata, see Steven Jan, "Using Galant Schemata as Evidence for Universal Darwinism", *Interdisciplinary Science Reviews* 38/2 (2013), 149–168. On the remaining terms frame (Minsky), prototype (Rosch), idealtype/archetype (Meyer), and model (Stachowiak), see Gjerdingen 1988 and 2007.

rently become almost a synonym of standardized *voice-leading* (or contrapuntal) patterns[12]; and this is how we also employ the term in this paper.

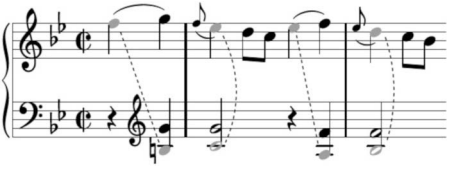
Voice-leading schemata are usually defined as configurations of two or more voices that move together through a sequence of stages, forming specific patterns of successive vertical intervals that occur within a specific tonal context.

Consider the example of the Fonte (a concept introduced as early as the mid-eighteenth century): The Fonte is usually defined as a four-stage pattern involving at least two voices. The bass moves through the scale degrees #1-2-7-1 of a major scale, while the soprano follows the pattern 5-4-4-3, thus producing the following sequence of vertical intervals: tritone → minor third → tritone → major third (see Example 1 for an abstract schema representation). In actual compositions, the schema prototype can be elaborated in multiple ways. For instance, the notes belonging to one stage can be displaced in time, as long as their belonging to a particular stage is indisputable. Also, structural notes can be flexibly ornamented by inserting additional notes. A real-world example illustrating the surface realization of a Fonte is given in Example 2.

Example 1: An abstract representation of a Fonte schema. The arrows denote the temporal sequence of vertical intervals (with dashed lines connecting the scale degrees in the upper voice and the bass voice).

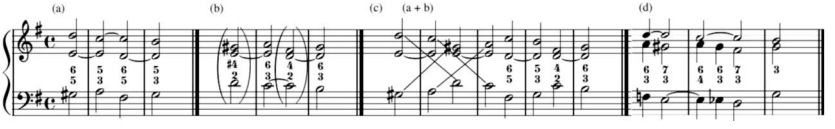


12 In German, these patterns are commonly referred to as 'Satzmodelle'; see, for instance, the special issue in the *Zeitschrift der Gesellschaft für Musiktheorie* 4/1–2 (2007), https://www.gmth.de/zeitschrift/ausgabe-4-1-2007/inhalt.aspx.

Example 2: A Fonte instance in Wolfgang Amadé Mozart's Piano Sonata in B-flat major, K 281, opening bars of final movement. The two diagonal dashed lines show note displacement within a stage. The notes occurring between the highlighted structural notes represent embellishments (or elaborations). (The figure is taken from Finkensiep et al. 2018.)

While the correct interval pattern is a central property of any schema instance, it is not a sufficient one. The selected notes must also provide the contrapuntal template for its context, such that all the notes contained in the time span covered by the schema instance can be meaningfully interpreted as ornamentations of the selected notes.

Although the Fonte seems to be a comparatively simple object, a variety of subcategories can be produced (and have in fact been used throughout music history) by applying standard operations such as voice shifting, voice exchange, note insertion, and manifold combinations thereof (see Example 3).

Example 3: The Fonte in its prototypical version (a) and variants produced by voice shifting (b), voice exchange (c), and note insertion (d).



Challenges of modelling and operationalization

To enable music theory and computer-assisted corpus research to efficiently interact, theories need to be sufficiently precise and formalizable, and the abstract concepts used need to be operationalized, that is, translated into observable entities. Since our project is concerned with identifying in digital corpora one particular class of musical objects such as voice-leading schemata, the very need to operationalize

and formalize these objects requires one to reconsider and refine traditional descriptions and definitions. Also, it may involve straightening out incompatible aspects between historical and modern definitions of particular schemata. The flipside of such strict definitions is that one may be forced to abstract from an object's fuzzy, but rich aspects in terms of contextual connotation. This is why operationalization is a notoriously intricate issue.

Even for such a well-known schema as the Fonte, several (slightly divergent) definitions exist in the music theoretical literature. Starting with the historical literature, Joseph Riepel (who coined the term Fonte) describes the schema as follows: "[…] gleich wie das *Monte* einen Ton hinauf – also steigt das *Fonte*, wie du weißt, einen Ton herab; das erste Glied wendet sich nämlich in die Secund *Tertz minor*, und das zweyte Glied in den Hauptton *Tertz major*".[13] Riepel thus emphasizes the sequential component: the first part as a whole is moved down by whole-step, thus producing a shift from minor to major (and therefore, in synaesthetic terms, from darkness to brightness).

A much broader definition is proposed by Markus Schwenkreis, who basically equates a Fonte with a descending-fifth sequence; as a result, there is no intrinsic extension of the schema in terms of the number of stages involved.[14] Finally, Robert Gjerdingen defines the Fonte primarily in terms of scale degrees in the soprano and bass: "The 7–1 ascent in the bass is matched by a 4–3 descent in the melody, often terminating a larger 6–5–4–3 descent […] All these features together suggest a Fonte prototype of four events arranged into two pairs […]."[15] In other words, the resulting vertical intervals are seen as a product of the contrapuntal combination of the two outer voices. This definition cannot capture those instances, however, in which no seventh (the local 4) is involved in the dominant (or V) chords launching each pair of events.

Describing schemata only in terms of a specific combination of voices and, therefore, as a succession of vertical intervals may be seen as insufficient, as schemata typically also feature a particular harmonic signature which can be expressed in terms of Roman numerals (which in turn assume chords, their roots, and their relation to a given key). The harmonic signature for the main variant of

13 Joseph Riepel, *Erläuterung der betrüglichen Tonordnung* (Augsburg 1765), 24 ('as the Monte goes up a step– so, as you know, the Fonte goes down a step; namely, the first member turns to the minor-mode second scale degree, and the second member to the main key tonic' (translation by the authors)).

14 Markus Schwenkreis (ed.), *Compendium Improvisation – Fantasieren nach historischen Quellen des 17. und 18. Jahrhunderts* (Basel: Schwabe, 2018), 87. A similar definition can be found in Hubert Mossburger, *Ästhetische Harmonielehre: Quellen, Analysen, Aufgaben*, vol. II (Wilhelmshaven: Noetzel, 2012), 770ff.

15 Gjerdingen, *Music in the Galant Style*.

the Fonte, for instance, would read as V65/ii–ii–V65–I. This harmonic layout is sometimes seen as a mere by-product of the voice combination described above.

Overall, the schema descriptions found in the scholarly literature are often somewhat imprecise and leave room for interpretation, which means they cannot readily be used for the purposes of computer-assisted music analysis. Therefore, in our project we had to come up with a (formal) model that translated the ideas described above into observable entities.

Generally, a formal model is a description of a particular segment of the world in terms of *entities* and their *relationships*, using the language of mathematics.[16] Since models aim to specify which aspects of the world are included and which are to be ignored, they inevitably involve abstraction from and simplification of the world under study.[17] As long as the researcher keeps in mind that the model is not to be confused with the richer real-world object, the model-based approach comes with the obvious advantage that otherwise diverse objects can now be compared with one another, thus enabling more general insights (see Conclusion).

From the perspective of model-based music theory, voice-leading schemata define the relationships between such entities as structural (schematic) and ornamental (non-schematic) notes and, in so doing, generalize over a wide variety of note configurations on the musical surface. In other words, schematic models provide an 'explanation' of the musical surface. The determination of how exactly the schematic core notes are elaborated in approaching the musical surface is the task of a comprehensive theory of counterpoint and historical styles, which has yet to be developed, though.

Our model of schemata

A notionally *comprehensive* model of musical schemata would no doubt require including a wealth of structural information about keys, harmonies, metrical weights, and distinct voices. In our model, however, we do not make any assumptions about these components in order not to considerably increase the complexity of the model.

As for the voices involved in the actual polyphonic structure of a piece, we do not make any assumption about their number and nature. Also, we do not aim to define

16 Christoph Finkensiep, Markus Neuwirth and Martin Rohrmeier, "Music Theory and Model-Driven Corpus Research", in: Daniel Shanahan, Ashley Burgoyne and Ian Quinn (eds.): *Oxford Handbook of Music and Corpus Studies* (New York: Oxford University Press, forthcoming). Generally on model-based digital humanities, see Julia Flanders and Fotis Jannidis (eds.), *The Shape of Data in Digital Humanities: Modeling Texts and Text-based Resources* (London: Routledge, 2018).

17 In contrast to 'approximation', which trades a simpler model for slight inconsistencies with the real world, 'abstraction' does not introduce any world-model contradictions.

rules that would help us to determine which note events on the musical surface belong to which voices.[18] Further, note events are not assigned scale-degrees within a particular key context, as this would require an analysis of the musical input in terms of tonal keys; nor does our model assign metrical weights to schema events. Instead, we opt for an underspecified and somewhat simplified model: voice-leading schemata are modelled merely as a specific succession of vertical intervals—a series of 'stages' based on a fixed number of voices. Each vertical interval in the model structure constitutes a 'stage' that contains one note per voice; and the number of stages is defined as fixed for each schema type.[19] Further, we assume that on the musical surface the structural notes belonging to a stage may be displaced in time, without making any restrictions on the number of intervening note events. Heuristically, however, we define a temporal limit where the distance is constrained between the displaced note events forming a stage.

The prototype for each schema variant (or subtype) is specified by using a formal notation. For instance, the prototype of the two-voice Fonte is encoded as: 'fonte.2': [['a1', 'P5'], ['M2', 'P4'], ['M7', 'P4'], ['P1', 'M3']], where '2' indicates the two-voice variant of the Fonte. Each note is given as an interval to some arbitrary reference point (for instance, P5 stands for 'perfect fifth', M2 for major second etc.). Since this comprises all possible transpositions of the schema, it is not necessary to know the reference key.

The challenge of finding schemata by using gram-based methods

The fact that (polyphonic) music consists of multiple simultaneous voices, and that the schematic core tones can be greatly embellished, makes it challenging to find voice-leading schemata in digitally encoded corpora. (Local) gram-based methods, which are standard in fields such as computational linguistics, constituted the first option that we considered. Gram-based methods extract short sequences from a longer stream of entities (e.g., letters or words in language; notes or chords in music). The most basic gram model, the n-gram, is just a *consecutive* subsequence in the input stream that has n elements.

18 Voice separation is itself a demanding theoretical and computational problem; see, for instance, Tillman Weyde and Reinier de Valk, "Chord- and Note-based Approaches to Voice Separation", in: David Meredith (ed.): *Computational Music Analysis* (Cham: Springer, 2016), 137–154.

19 This is another simplification: schemata such as the 5–6 progression or the descending-fifths sequence do not have a fixed size in terms of the number of stages involved.

While n-gram approaches are useful for uncovering contrapuntal patterns that consist of elements appearing adjacent on the musical surface,[20] skip-gram approaches allow one to capture non-adjacent structural elements.[21] By extending the n-gram idea, skip-grams produce subsequences by leaving out (or 'skipping') up to k elements.[22] Thus, skip-grams may reveal particular patterns that are otherwise 'disguised' by intervening events.

However, using a conventional skip-gram approach did not prove sufficient for our purpose. Both n-grams and their skip-gram extensions assume the distance between entities to be determined by their positions in the stream. While this assumption might be reasonable for text, monophonic melodies, or chord sequences, it is problematic for other applications that involve multiple (simultaneous) temporal streams, in particular streams of musical events such as notes. Therefore, it is desirable to measure the distance between events (notes) based on their *timing* information (that is onset, offset, and duration). Second, while notes might be simultaneous in a score, they occur sequentially in a stream or list-of-notes representation, which becomes problematic if distance is measured by index.

As a novel approach to the problem of finding polyphonic patterns in music, we developed an algorithm that extended previous n-gram and skip-gram approaches by using an arbitrary distance measure (e.g., temporal distance).[23] This allows skip-grams to be applied to non-sequential structures such as polyphonic music. Furthermore, the ability to operate on individual notes enables us to group non-simultaneous notes into stages. The stages can then be grouped into schema candidates in a second pass. Note, however, that due to the exhaustive search and a great number of possible note combinations, our resulting dataset is extremely unbalanced, containing many more accidental occurrences of the pre-defined interval pattern (false positives) than true schema instances. To further reduce the number of schema candidates, guided by prior music-theoretical intuition, we restricted the window size

20 E.g., Mathieu Bergeron and Darrell Conklin, "Subsumption of Vertical Viewpoint Patterns", in *International Conference on Mathematics and Computation in Music* (Springer: Berlin, Heidelberg, 2011), 1–12; and Christopher Antila and Julie Cumming, "The VIS Framework. Analyzing Counterpoint in Large Datasets", in: Hsin-Min Wang, Yi-Hsuan Yang and Jin Ha Lee (eds.) *Proceedings of the 15th International Society for Music Information Retrieval Conference* ISMIR (Taipei, Taiwan, 2014), 71–76.

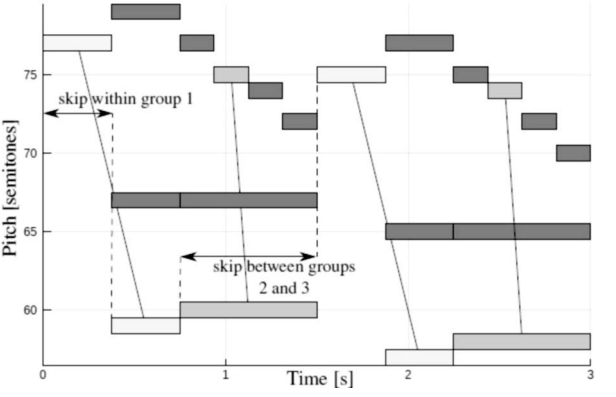
21 David R. Sears, Andreas Arzt, Harald Frostel, Reinhard Sonnleitner and Gerhard Widmer, "Modeling Harmony with Skip-Grams", in: *Proceedings of the International Society on Music Information Retrieval* (2017), 332–338.

22 D. Guthrie, B. Allison, W. Liu, L. Guthrie and Y. Wilks, "A Closer Look at Skip-Gram Modelling", in: *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC'06)* (Genoa: European Language Resources Association (ELRA), 2006), 1222–1225.

23 See Finkensiep et al., "Generalized Skipgrams for Pattern Discovery"; see also Sears and Widmer, "Beneath (or Beyond) the Surface".

to a maximal note displacement of one bar per stage and a maximal distance of one bar between the onsets of two adjacent stages (for an example of schema-matching using our skip-gram approach, see Example 4).

Example 4: The figure shows an example of the application of skip-grams to polyphonic music displayed in piano-roll visualization. The highlighted notes are members of the skip-gram; the stages are indicated by solid lines between notes belonging to the same stage. The skip-gram pattern refers to the schema instance shown in Example 2. (The figure is taken from Finkensiep et al. 2018.).



The dataset

The data created in the project and used for schema matching consist of digital musical scores and expert-labelled schemata added to them.[24] Our dataset is based on the full set of Mozart's piano sonatas encoded in MusicXML format. These 18 sonatas with three movements each (thus 54 movements in total) were composed between 1774 and 1789 and constitute a prominent sample of the classical style. The pieces in the dataset contain 103,829 notes in total distributed over 7,500 measures, with 244 hand-annotated true schema instances. They are complemented by 190,994 false instances (99.87% vs. 0.13% true instances) which were additionally found by the skip-gram matcher for the selected schema types and subtypes. In Finkensiep et al.

24 The underlying annotation standard has been developed in the course of this project. The dataset is stored for version control in a GitHub repository. For historical sources of classical music, there are no copyright issues. See Finkensiep et al., "Voice-Leading Schema Recognition".

2020, we selected 10 schema types and 20 subtypes from a comprehensive lexicon of schemata. Keeping the high combinatorial complexity of multi-voice structures in mind, we decided to restrict this study to two-voice schema variants. Furthermore, we reduced the number of candidates to at most 25 per group of temporally overlapping candidates.

Schema instances are encoded as nested arrays of notes in the same form as the corresponding prototypes. Instances may deviate from the shape of the prototype if (a) a note that would repeat its predecessor (e.g. the second 4 in the Fonte) is held over or missing, or (b) two adjacent voices merge and are represented by a single note on the surface.

The schema annotation was performed in a hybrid fashion. First, we developed our skip-gram-based schema matcher that located potential instances of schemata in a digital score and helped the user visualize and explore these candidates.[25] Second, we developed a schema annotation app that allowed users to annotate musical schemata. The app interacts with the schema matcher so that algorithmically computed schema instances can be taken into account or modified in the human annotation process.

Regarding the human-machine interaction, we observe an effect of mutual interaction: the human expert assesses schema instances as classified by the algorithm while the algorithm also identifies schema instances which might surprise the expert or which the expert has missed. We begin to observe a phenomenon that was previously seen in other developments of Artificial Intelligence: as evident in the research on games such as chess, backgammon, poker, or Go from the 80s onwards, human expertise is not the only infallible source of evidence and computational models (begin to) infer patterns that may partly extend expert knowledge.[26]

Scrutinizing domain expert intuitions

The dataset compiled by computer-assisted annotation has been explored by using statistical methods. The results help scrutinize theoretical intuitions and hypotheses about the use of individual schemata and their frequency of occurrence.

Since the aim of our project was not to solve an engineering task, we chose to design a formal model of schemata that explicitly reflected their structural properties

25 Note that schema visualization is an integral part of schema matching and annotation apps.

26 On AI and music as well as related challenges, see, for instance, Martin Rohrmeier, "On Creativity, Music's AI Completeness, and Four Challenges for Artificial Musical Creativity", in: *Transactions of the International Society for Music Information Retrieval* 5/1 (2022).

(rather than a neural network, for instance). This way, the results and model properties remain interpretable for the human theorist and also provide feedback for music theory by usefully defining schemata and its related set of features. The very process of developing the different versions of the computational models (leading to the current, but not final, model) is itself a rich resource, where more traditional music analysis and formal modelling enrich each other in various ways (see Conclusion).[27]

While classifying musical building blocks as instances of an underlying schema, human listeners and analysts rely, often unconsciously, on a variety of features. Whether or not this is done in a consistent fashion is open to debate and empirical scrutiny. In our project, the uncertainty associated with schema identification and classification played an important role. There are numerous borderline cases which do not neatly fit into one of the available categories, or which combine elements from two or more categories. In this regard, computational approaches prove powerful in testing the criteria on which human judgment implicitly relies. In our case of schema detection, these (pitch- and rhythm-related) criteria are 'complexity', 'regularity', and 'salience'. While complexity involves the issue of whether two structural events of one stage occur at the same time or at different onsets, regularity concerns, for instance, the rhythmic similarity across stages;[28] and salience is related to duration and metrical weight.

In Finkensiep et al. (2020), we have evaluated the above features for schema recognition by using a binary classifier. Here, we have shown that parallelism (i.e. rhythmic regularity aligned to the metrical grid) has indeed a strong positive influence, thus indicating a preference for regular temporal organization. Generally, true schema instances exhibit higher internal regularity and lower complexity (e.g. in terms of note displacement) than non-instances do. In contrast, properties related to intrinsic salience (e.g. duration or metric weight of the matched notes) are less important.

As the results in Finkensiep et al. (2020) show, distinguishing between incidental and structural note configurations based on a small number of musically and cognitively motivated heuristics works well in the vast majority of cases. Even if a number of misclassifications remain, a closer look at these cases provides valuable insights into the problem at hand. First, the main limitation of our approach is that the model

27 The proximity of formal modelling in computer science to hermeneutic techniques as practised in music analysis is discussed in Michael Piotrowski and Markus Neuwirth (2020), "Prospects for Computational Hermeneutics", in: Cristina Marras and Marco Passarotti (eds.): *Proceedings of the 9th Annual Conference of the AIUCD* (Milano, Jan. 15–17, 2020), Associazione per l'Informatica Umanistica e la Cultura Digitale (AIUCD), 204–209.

28 See, for instance, Mathieu Giraud, Ken Déguernel and Emilios Cambouropoulos, "Fragmentations with Pitch, Rhythm and Parallelism Constraints for Variation Matching", in: *International Symposium on Computer Music Multidisciplinary Research (CMMR)*, 2013, 298–312.

assesses suggested schema instances individually without considering or comparing them to alternative interpretations. In many cases, the main reason for human experts to reject a candidate does not seem to be a lack of plausibility of the match itself, but the unavailability of a 'better explanation', that is, an alternative analysis of the context of the match that identifies a more plausible contrapuntal scaffold.

A second insight concerns the idea of schema itself and its relation to a classification task. From a cognitive perspective, a schema does not need to be instantiated unambiguously or even completely. It is sufficient if listeners recognize the schema as the template for the surface events, or if they understand the composer's intention to evoke the schema. In this regard, discrete binary classification into instances and non-instances may be as unattainable as it is undesirable, falling short of the complexity that the relationship between schema and realization can exhibit.

Conclusion: The interactive potential between close and distant reading in music research

By way of conclusion, we would like to consider the relationship between traditional music theory (as a humanities discipline) and digital/quantitative research on a more general level, pondering the various objections raised against their interaction.

The frequently voiced accusation of positivism (in its various forms) that the generalizing impulse of quantitative research is not capable of capturing the 'qualities' realized in each individual (historical) case is often a reflexive reaction to corpus-based computational studies. The widespread scepticism towards quantitative approaches in historical or cultural music research is based on a number of concerns.

Firstly, one may argue that historically differentiating, context-sensitive 'close reading' and (much) more coarse-grained 'distant reading' are mutually exclusive modes of research. The observation of patterns in a collection of objects by necessity requires abstraction from details in order to make the objects comparable (see above on modelling and its challenges). The individuality-driven analysis as practised in hermeneutically oriented studies has a different purpose and research interest than the observation of common features or trends in a large corpus.

Secondly, qualitative hermeneutic methods usually operate with a larger number of categories; and they are furthermore characterized by a higher degree of granularity in the sense of a 'thick description' (Clifford Geertz) than the ideally few and more large-grained categories of corpus research. This is precisely why a corpus study cannot be obtained from the accumulation of many small hermeneutic analyses.

Thirdly, there is the fear that in the long run computational and quantitative methods would eventually replace qualitative hermeneutic approaches and make human analysts expendable.

Nonetheless, without casting doubt on the principle validity of points 1 and 2, we want to argue that close and distant readings can be bridged such that they complement each other in productive ways. To begin with, qualitative insights—or pre-empirical intuitions—serve as a starting point that allows us to come up with musically and historically sensitive formal definitions of our research objects (i.e., schemata); and hypotheses concerning their historical use are to be tested on a broader empirical basis. Furthermore, hermeneutics and computer-assisted music analysis share a common concern for uncertainty which arises at all stages of the research agenda including data production, analysis, and interpretation. In our project, uncertainty arises in terms of schema identification and classification.

Finally, 'distant reading' has the potential of providing important insights on which detailed individual case studies can additionally be based. The digital representation of musical objects allows scholars to systematically scrutinize and compare different degrees of granularity and their information content in relation to some prior hypotheses and research aims. Qualitative hermeneutic questions can thus be raised to a new scientific level by corpus research without losing its own legitimacy.

More generally, we are currently in the midst of a digital revolution that may prove to be a major turning point in human society. The amount of music available and the historical scale that can now be explored are much higher than in previous, more traditional research. Computational modelling brings musicological research to an entirely new and unprecedented level. But computational musicology, being still in its early phase of development (despite the use of electronic and digital tools for music for a long time), has yet to unfold its enormous potential. At present, one of the biggest challenges in the field is the synthesis between traditional, analytical, and historical methods, not to mention the new methods and tools coming from digital research in order to reshape the discipline of musicology for the future. Exploration of the many possibilities of interaction between quantitative and qualitative methods in an integrative 'mixed methods' framework remains one of the urgent tasks of future research.

Bibliography

Allegraud, Pierre, Louis Bigo, Laurent Feisthauer, Mathieu Giraud, Richard Groult, Emmanuel Leguy and Florence Levé. "Learning Sonata Form Structure on Mozart's String Quartets", In *Transactions of the International Society for Music Information Retrieval* 2, 1 (2019): 82–96.

Antila, Christopher and Julie Cumming. "The VIS Framework. Analyzing Counterpoint in Large Datasets". In *Proceedings of the 15th International Society for Music Information Retrieval Conference*, ISMIR, edited by Hsin-Min Wang, Yi-Hsuan Yang and Jin Ha Lee, 71–76, Taipei, Taiwan, 2014.

Bergeron, Mathieu and Darrell Conklin. "Subsumption of Vertical Viewpoint Patterns". In *International Conference on Mathematics and Computation in Music*, 1–12, Springer: Berlin/Heidelberg, 2011.

Byros, Vasili. "Towards an 'Archaeology' of Hearing: Schemata and Eighteenth-century Consciousness". In *Musica Humana* 1, 2 (2009): 235–306.

Byros, Vasili. "Trazom's Wit: Communicative Strategies in a 'Popular' Yet 'Difficult' Sonata". In *Eighteenth-Century Music* 10, 2 (2013): 213–252.

Caplin, William E. "Topics and Formal Functions: The Case of the Lament". In *The Oxford Handbook of Topic Theory*, edited by Danuta Mirka, 415–452. Oxford: Oxford University Press, 2014.

Conklin, Daniel and Mathieu Bergeron. "Discovery of Contrapuntal Patterns". In *Proceedings of the International Society on Music Information Retrieval*, 2010, 201–206.

Dittrich, Marie-Agnes. "'Teufelsmühle' und 'Omnibus'". In *Zeitschrift der Gesellschaft für Musiktheorie* 4, 1–2 (2007): 107–121. https://doi.org/10.31751/247

Duane, Ben. "Melodic Patterns and Tonal Cadences: Bayesian Learning of Cadential Categories from Contrapuntal Information". Journal of New Music Research 48, 3 (2019): 197–216.

Finkensiep, Christoph, Markus Neuwirth and Martin Rohrmeier 2018. "Generalized Skipgrams for Pattern Discovery in Polyphonic Streams". In *Proceedings of the 19th International Society for Music Information Retrieval Conference* (ISMIR), Paris 2018, edited by Emmanouil Benetos, Emilia Gómez, Xiao Hu and Eric Humphrey, 547–553.

Finkensiep, Christoph, Markus Neuwirth and Martin Rohrmeier. "Music Theory and Model-Driven Corpus Research". In *Oxford Handbook of Music and Corpus Studies*, edited by Daniel Shanahan, Ashley Burgoyne and Ian Quinn, New York: Oxford University Press. (forthcoming)

Finkensiep, Christoph, Ken Déguernel, Markus Neuwirth and Martin Rohrmeier. "Voice-Leading Schema Recognition Using Rhythm and Pitch Features". In *Proceedings of the International Society for Music Information Retrieval* (ISMIR), Montreal 2020, edited by Julie Cumming, Jin Ha Lee, Brian McFee et al., 520–526.

Flanders, Julia and Fotis Jannidis (eds.). *The Shape of Data in Digital Humanities: Modeling Texts and Text-based Resources*. London: Routledge, 2018.

Giraud, Mathieu, Ken Déguernel and Emilios Cambouropoulos. "Fragmentations with Pitch, Rhythm and Parallelism Constraints for Variation Matching". In *International Symposium on Computer Music Multidisciplinary Research* (CMMR), 2013, 298–312.

Gjerdingen, Robert O. *Music in the Galant Style*, New York: Oxford University Press, 2007.

Hentschel, Johannes, Markus Neuwirth and Martin Rohrmeier. "The Annotated Mozart Sonatas: Score, Harmony and Cadence". In *Transactions of the International Society for Music Information Retrieval 4*, 1 (2021): 67–80. http://doi.org/10.5334/tismir.63.

Jan, Steven. "Using Galant Schemata as Evidence for Universal Darwinism". In *Interdisciplinary Science Reviews 38*, 2 (2013): 149–168.

Katsiavalos, Andreas, Tom Collins and Bret Battey. "An Initial Computational Model for Musical Schemata Theory". In *Proceedings of the International Society on Music Information Retrieval*, 2019, 166–172.

Meredith, David, Kjell Lemström and Geraint A. Wiggins. "Algorithms for Discovering Repeated Patterns in Multidimensional Representations of Polyphonic Music". In *Journal of New Music Research 31*, 4 (2002): 321–345. DOI:10.1076/jnmr.31.4.321.14162.

Meyer, Leonard B. "Innovation, Choice and the History of Music". In *Critical Inquiry 9*, 3 (1983): 517–544.

Moss, Fabian C., Markus Neuwirth, Daniel Harasim and Martin Rohrmeier. "Statistical Characteristics of Tonal Harmony: A Corpus Study of Beethoven's String Quartets". In *PLoS ONE 14*, 6 (2019), e0217242.

Mossburger, Hubert. Ästhetische Harmonielehre: Quellen, Analysen, Aufgaben, vol. II, Wilhelmshaven: Noetzel, 2012.

Neuwirth, Markus and Martin Rohrmeier. "Wie wissenschaftlich muss Musiktheorie sein? Chancen und Herausforderungen musikalischer Korpusforschung". In *Zeitschrift der Gesellschaft für Musiktheorie 13*, 2 (2016): 171–193.

Piotrowski, Michael and Markus Neuwirth (2020). "Prospects for Computational Hermeneutics". In *Proceedings of the 9th Annual Conference of the AIUCD* (Milano, Jan. 15–17, 2020), edited by Cristina Marras and Marco Passarotti, Associazione per l'Informatica Umanistica e la Cultura Digitale (AIUCD), 204–209.

Rice, John. "The Morte: A Galant Voice-Leading Schema as Emblem of Lament and Compositional Building-Block". In *Eighteenth-Century Music 12*, 2 (2015): 157–181.

Rohrmeier, Martin. "On Creativity, Music's AI Completeness, and Four Challenges for Artificial Musical Creativity". In *Transactions of the International Society for Music Information Retrieval 5*, 1 (2022): 50–66. DOI: http://doi.org/10.5334/tismir.104

Rumelhart, David E. "Schemata: The Building Blocks of Cognition". In *Theoretical Issues in Reading Comprehension*, edited by Rand J. Spiro, Bertram C. Bruce and William F. Brewer, 33–58, Hillsdale, NJ: Lawrence Erlbaum, 1980.

Schank, Roger and Robert P. Abelson. *Scripts, Plans, Goals, and Understanding: An Inquiry into Human Knowledge Structures*, New York: Halsted, 1977.

Schwenkreis, Markus (ed.), *Compendium Improvisation—Fantasieren nach historischen Quellen des 17. und 18. Jahrhunderts*, Basel: Schwabe, 2018.

Sears, David R., Andreas Arzt, Harald Frostel, Reinhard Sonnleitner and Gerhard Widmer. "Modeling Harmony with Skip-Grams". In *Proceedings of the International Society on Music Information Retrieval*, 2017, 332–338.

Sears, David R., Marcus T. Pearce, William E. Caplin and Steven McAdams. "Simulating Melodic and Harmonic Expectations for Tonal Cadences Using Probabilistic Models". In *Journal of New Music Research* 47, 1 (2018): 29–52.

Sears, David R. and Gerhard Widmer. "Beneath (or Beyond) the Surface: Discovering Voice-leading Patterns with Skip-grams". In *Journal of Mathematics and Music 15*, 3 (2020): 1–26. DOI: 10.1080/17459737.2020.1785568.

Symons, James. "A Cognitively Inspired Method for the Statistical Analysis of Eighteenth-Century Music, as Applied in Two Corpus Studies". PhD Thesis, Northwestern University, 2017.

Telesco, Paula J. "Enharmonicism and the Omnibus Progression in Classical-Era Music". In *Music Theory Spectrum* 20, 2 (1998): 242–279.

Weiß, Christof, Frank Zalkow, Vlora Arifi-Müller, Meinard Müller, Hendrik Vincent Koops, Anja Volk and Harald G. Grohganz. "Schubert Winterreise Dataset: A Multimodal Scenario for Music Analysis". In *ACM Journal on Computing and Cultural Heritage (JOCCH) 14*, no. 2 (2021): 1–18.

Weyde Tillman and Reinier de Valk. "Chord- and Note-based Approaches to Voice Separation". In *Computational Music Analysis*, edited by David Meredith, Cham: Springer, 2016, 137–154.

#MODELLING

The term #Modelling, as understood by the Oxford English Dictionary, primarily denotes the creation of representation, or description, of structure. In many cases of everyday communication, this refers to art, crafts or engineering or to abstract processes in science in which visualization and abstraction are of relevance. *

As such, the term seems to parallel #Visualization. Yet, a modelling tool is not only "a tool used for modelling or working a material" but "a program which assists with the manipulation of data for the construction of a computer model" as well. Interestingly, the term computer model is listed only as a compound of computer in "the sense 'carried out or created by means of a computer or computers'" without further explanation or contextualization.

It seems that this definition does not do justice to the actual role of #Modelling in computer science and digital humanities. The terminology emerged some decades ago as did many of the computer science and digital humanities terms. Still it is not clear whether the sheer complexity and importance of the phenomenon (yet) prevents a definition suitable for everyday use, or whether it is discursive as a technical term to such a degree that a hermetic discourse within computer science takes place and hinders agreement at this time.

The definitions by the mixed methods project teams sketch an overall impression of meaning that remains fully within the realms of the digital. Although only hinted at, the models seem to do in fact the practical work of bridging the gap between humanities and the digital: "The challenge (…) is to come up with modelling decisions that are computationally feasible and still faithful to the object under study" (Bach-Beatles). Beyond this information, however, the actual characteristic of such modelling process remains vague. Definitions and project descriptions hint at models positioned downstream of #Quantification enabling different kinds of #Machine Learning. Beyond that, the information is indirect, stating "parameters of a computational model are tuned based on data (…) that have already been classified into a range of classes" for example, or that there are attention models, thus suggesting a broader range of concepts for models (Rhythmicalyzer).

In sum, the term #Modelling describes either a very fluid concept within computer science/digital humanities or such a basic element that a precise definition seems unnecessary. It is needed, however, to foster communication during the development of digital humanities and mixed methods approaches.

* "Modelling | modeling, n.". in: *Oxford English Dictionary (OED)*, Third Edition, September 2002; most recently modified version published online March 2022, https://www.oed.com/ [accessed 25.10.2022].

Title: Rhythmicalizer. A digital tool to identify free verse prosody (www.rhythmicalizer.net)
Team: Burkhard Meyer-Sickendiek, Timo Baumann, Hussein Hussein
Corpus: German spoken poetry from Lyrikline website (www.lyrikline.org)
Field of Study: Poetry analysis, digital humanities, supervised machine learning
Institution: Free University of Berlin, Department of Literary Studies, Berlin, Germany; University of Hamburg, Department of Informatics, Hamburg, Germany
Methods: free verse prosody, machine learning, classification of rhythmical patterns
Tools: Text-Speech Alignment, Part-of-speech tagging of written text, Wavesurfer/Snack (speech annotation, visualization, and analysis), DyNet (neural network learning), WEKA (various ML techniques)
Technology: machine learning for classification of rhythmical patterns

Free Verse Prosodies: Identifying and Classifying Spoken Poetry Using Literary and Computational Perspectives (Rhythmicalizer)

Timo Baumann, Hussein Hussein, Burkhard Meyer-Sickendiek

Abstract *At least 80% of modern and postmodern poems exhibit neither rhyme nor metrical schemes such as iamb or trochee. However, does this mean that they are free of any rhythmical features? The US American research on free verse prosody claims the opposite: Modern poets like Whitman, the Imagists, the Beat poets and contemporary Slam poets have developed a post-metrical idea of prosody, using rhythmical features of everyday language, prose, and musical styles like Jazz or Hip Hop. It has spawned a large and complex variety in their poetic prosodies which, however, appear to be much harder to quantify and regularize than traditional patterns. In our project, we examine the largest portal for spoken poetry Lyrikline and analysed and classified such rhythmical patterns by using pattern recognition and classification techniques. We integrate a human-in-the-loop approach in which we interleave manual annotation with computational modelling and data-based analysis. Our results are integrated into the website of Lyrikline. Our follow-up project makes our research results available to a wider audience, in particular to high school-level teaching.*

Introduction

The theoretical starting point of our work is an American research discussion that has not previously been noted in German studies: the so-called free verse prosody. The most important theoretical orientation of this new theory of poetry was provided in 1980 by Charles O. Hartman in his influential study *Free Verse: An Essay on Prosody*. Hartman assumed that the free verse was prosodical, but not metrical—"the prosody of free verse is rhythmic organization by other than numerical modes"[1]—and therefore defined the prosody of the free verse via the term rhythm. Prosody was therefore a "system of rhythmic organization that governs the con-

1 Charles O. Hartmann: *Free Verse: An Essay on Prosody* (Princeton: University Press, 1980) 14.

struction and reading of a poem"[2]. Based on these theses, a discussion about the design principles of this free verse prosody has been developed in the US-American research area, reaching right down to the present day.

As an example, we offer the most important poem demonstrating the 'breath controlled line' in German poetry: Ernst Jandl's poem 'beschreibung eines gedichtes' (description of a poem) taken from the volume *der gelbe hund*:

bei geschlossenen lippen
ohne bewegung in mund und kehle
jedes einatmen und ausatmen
mit dem satz begleiten
langsam und ohne stimme gedacht
ich liebe dich
so daß jedes einziehen der luft durch die nase
sich deckt mit diesem satz
jedes ausstoßen der luft durch die nase
und das ruhige sich heben
und senken der brust[3]

The poem shown in its textual form is available as spoken by the author via the QR code printed above and can also be accessed via the URL https://www.lyrikline.org/de/gedichte/beschreibung-eines-gedichtes-1233. Even through reading the text one

2 Ibid.

3 Ernst Jandl: *der gelbe hund* (Darmstadt/Neuwied: Luchterhand, 1985).

will notice that the spoken delivery of the poem plays a vital function in its perceptive impression. It is this focus on *spoken poetry* that we focus on in our work.

Digitalization has made our field of research—the genre of audio poetry—possible in the first place. Of course, printed poems have been around for centuries, but audio poems have only existed at the present quantity and frequency since about 2000 with the development of the Internet. Reference has been made to the changed conditions of production and reception that prevail in the audio production of poetry in the digital age. While poetry readings change poems as text form in a lasting way, since they take place in the contingent space of the public in the age of the Internet, poetry can also be performed and intoned as a silent reading. According to Ursula Geitner, with the "end of public and domestic declamation practice around 1900", the "eloquence of the body is also reduced", insofar as this "new reading art […] concentrates on the author's voice."[4]

The idea of the project that we report in this chapter is twofold: we introduce the discussion on free verse prosody in post-modern poetry into German literary studies and examine the theses of this discussion. We also support them by using data-based methods not only in the literary analysis, but also in the computational models which we introduce and which leverage machine learning to foster our understanding of the relevant classes and their interrelation.

The remainder of the chapter is structured as follows: we introduce Rhythmicalizer, the project in which we have undertaken our research, in the following section and describe the literary theory foundation, the data and the methods that we have used. We describe our primary literary and computational results within the discussion of our methods, as these were produced in a cyclical, iterative manner by the interplay of literary analysis and computational modelling. We then discuss our overall learning that goes beyond the immediate results in our conclusion.

Project Rhythmicalizer

In this section, we describe the goal of the project for classiying poems according to their rhythmical features and the partners in the project.

The aim of the project *Rhythmicalizer* (www.rhythmicalizer.net) is twofold: we test and verify the basic patterns of free verse prosody theory, which in the US and for English is a highly recognized scientific theory but neglected as a lyrical rhythm for German, within the framework of a digital corpus analysis. It is complicated by the nature of free verse prosody which is an almost purely *spoken* aspect of a poem and is

4 Geitner, Ursula: *Die Sprache der Verstellung. Studien zum rhetorischen und anthropologischen Wissen im 17. und 18. Jahrhundert* (Berlin: De Gruyter, 1992), 342. Translation: B.M.S.

not readily observable from the pure textual form as in traditional metric schemes. We therefore base our analyses on the audible form of the poem (instead of, e.g., attempting to derive such features in a generic way from the written form). We then group poems into *classes* that express different kinds of free verse prosody.

In the computational part of the project, we integrate the two modalities—speech and text—and build an automatic classification system that uses quantitative features to yield our classifications. Our system needs to overcome the issue of *data sparsity* (i.e., the relatively small number of poems available as compared to the need for large amounts for data commonly used by modern machine-learning techniques).

Finally, we are concerned with the matter of preprocessing the poems in our collection so that they can be classified in an automatic way; and ultimately, we attempt to gather additional, new insight from the automatic classification by using a *human-in-the-loop* approach, in which we interleave manual annotation with computational modelling and data-based analysis.

Our research is performed in collaboration with the internet portal '*Lyrikline*' (www.lyrikline.org), located in Berlin and initiated by Literaturwerkstatt Berlin, probably the most important internet portal for international poetry readings worldwide hosting contemporary international poetry as audio (read by the authors) and text (original versions & translations). It provides the melodies, sounds and rhythms of international poetry, read by the original authors. Users can listen to the poet and read the poems both in their original language and in various translations.

An important aspect to the realization of the project was a change in German copyright laws in March 2018 regarding the use of publicly available data for research purposes.[5] It allowed us to freely download and use the data, but hardly did it simplify collaboration with *Lyrikline*.

Corpus Development

The database of our research project is based on the internet portal *Lyrikline*. The digital material on *Lyrikline* contains more than 10,000 poems by over 1200 international poets from almost 80 different countries. Nearly 80% of them are post-metrical poems. We examine German and English poems, with a total of 392 poets (227 German, 165 English) and about 3840 poems (2465 German, 1376 English). We

5 Burkhard Meyer-Sickendiek, Hussein Hussein: "The New German Copyright Law for Science and Education (UrhWissG): Consequences for DH-Projects Working with Non-Academic Partners", in: *Proceedings of the 15th Conference on Knowledge Organization WissOrg'17 of the German Chapter of the International Society for Knowledge Organization (ISKO) 2017*, Berlin, December 2017.

use this corpus to determine rhythmic patterns, extended with cornerstone poems by some authors that are not available on *Lyrikline* but on other platforms, such as YouTube.

As part of our project, we built a tool[6] to systematically extract poems and their meta-data from *Lyrikline* in order to form a corpus. We then performed forced text-audio alignment (which we refined manually as necessary) in order to relate speech and text line by line. In addition, we used various tools to extract syntactic information from the text as required (such as parts of speech for each word). We needed to label the data in order to detect the classes in them (rhythmical patterns). For this reason, the philologist classified every poem after listening to their audio recording into one of our rhythmical patterns.

Of course, a primary difficulty in a project such as ours is the expert categorization of poems into classes as not all poems on *Lyrikline* fit into a category and not every exceptional poem merits its own category. This was why we defined a threshold value for poems to constitute a new pattern: in total we expected up to 25 different patterns and used a threshold of 1% of poems belonging to this pattern. Thus, with about 2400 German-language poems on *Lyrikline*, the threshold value was 25: about 25 poem samples should be found per pattern for it to be relevant. (Note that not all classes reach this level; also, we could not manually fully analyze all of *Lyrikline*; see also Section 4.3.).

Table 1 shows some key descriptive statistics of the poems as assigned to their classes. The full manual categorization of poems into classes is available for research purposes from the authors.

Table 1: Description of the data used in the experiment.

	Poems	Lines	Characters	Audio
Lyrikline (German sub corpus)	2392	61849	2025484	52 hours
Cadence	31	1079	34163	49 min
Parlando	34	1435	44323	67 min
Variable foot	34	878	23684	39 min
Hook style	36	1090	33178	48 min
Gestic rhythm	33	897	27741	44 min
Ellipsis	56	2154	62704	104 min

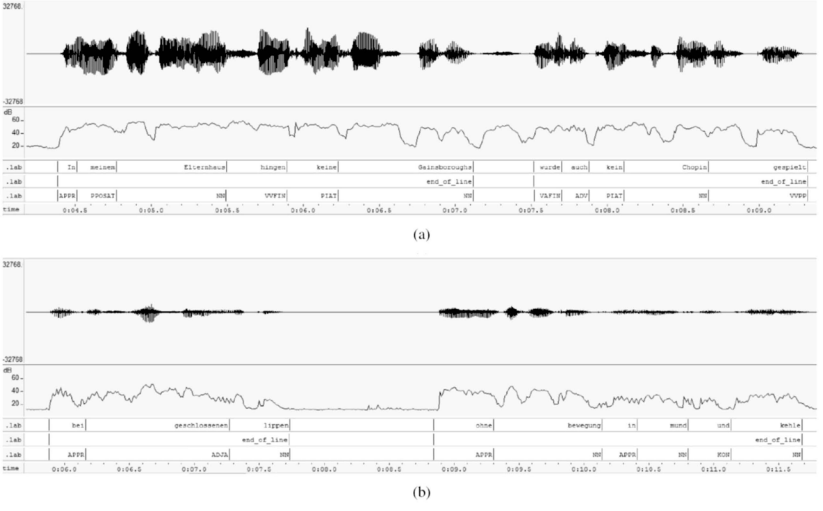
6 A program for extracting poems (text, audio and metadata) from www.lyrikline.org is available by request from the authors.

Permutation	30	1117	32041	46 min
Syllabic decomposition	21	540	12390	26 min
Lettristic decomposition	17	684	10007	31 min

Data Processing and Tools for Handling and Analysis

In order to simplify speedy manual analysis as well as hypothesis testing and to enable automatic processing, data on all poems are stored time-aligned, either in a word-by-word or in a line-by-line manner. Figure 1 shows an exemplary result of data preprocessing for two poems, each showing (from top to bottom) speech signal, intensity (dB), word alignment, end of line alignment, parser information (PoS-tagging), and time, by using the speech visualization tool Wavesurfer.[7]

Fig. 1: Poem analysis: (a) Analysis of the first two lines of the poem 'TEILS-TEILS' (English: Half Here, Half There) from the poet, Gottfried Benn, as an example of the 'Parlando' pattern. (b) As in (a) but for the poem 'beschreibung eines gedichtes' (English: Description of a Poem) from the poet, Ernst Jandl, as an example of the 'Variable Foot' pattern.



7 Kåre Sjölander, Jonas Beskow, "Wavesurfer–an open source speech tool", in: *Sixth International Conference on Spoken Language Processing*, 2000.

A forced alignment of text and speech for poems is performed using a text-speech aligner[8] which employs a variation of the Sail Align algorithm[9] implemented via Sphinx-4.[10] The line boundaries (the start of the first word and the end of the last word in each poetic line) are detected. The alignments are stored in a format that guarantees that the original text remains unchanged, which is important for recreating the exact white spacing in the poem (the white space is important as the text in a poem and helps readers know how to read a poem out). The forced alignment of text and audio in spoken poetry, especially in concrete and sound poetry, is non-trivial and often individual words or lines cannot be aligned. Therefore, the automatically extracted alignment information is manually corrected by the second author more than once (rectifying alignment information and in some cases correcting written text and audio files of poems) by listening to the audio file and looking at the waveform.

We also processed the text data of poems by using a statistical parser in order to extract syntactic features. The Stanford parser[11] is used to parse the written text of poems. The parser used the Stuttgart-Tübingen-Tag Set (STTS) table developed at the Institute for Natural Language Processing of the University of Stuttgart[12] for the parsing of German poems. A major problem in poem parsing is the absence of punctuation marks. In addition, many poems are written with special characters: sometimes the text is written in lower case with some words in upper case, which makes the recognition of sentence boundaries quite difficult. Furthermore, some sentences in the poems go beyond the line boundary and run on to the next line. Such unconnected syntactic elements result from the dissolution of poetic lines, caused by so-called enjambment. No countermeasure was taken by us to solve these problems since the correction of the problems was time-consuming and changed the poem form.

---

8 Timo, Baumann, Arne Köhn, Felix Hennig: "The Spoken Wikipedia Corpus Collection: Harvesting, alignment and an application to hyperlistening", *Language Resources and Evaluation*, vol. 52, no. 2, 303–329, 2018, special Issue representing significant contributions of LREC 2016.

9 Katsamanis, A., M. Black, P. G. Georgiou, L. Goldstein, and S. Narayanan, "Sail Align: Robust Long Speech-Text Alignment", in: *Proc. of Workshop on New Tools and Methods for Very-Large Scale Phonetics Research*, 2011.

10 Willie Walker et. Al., "Sphinx-4: A Flexible Open Source Framework for Speech Recognition", Mountain View, CA, USA, Tech. Rep., November 2004.

11 Anna Rafferty, Christopher D. Manning, "Parsing Three German Treebanks: Lexicalized and Unlexicalized Baselines", in: *Proceedings of the Workshop on Parsing German*, Stroudsburg, PA, USA: Association for Computational Linguistics, 2008.

12 Anne Schiller et. al., "Guidelines für das Tagging deutscher Textcorpora mit STTS (Kleines und größes Tagset)", available on http://www.sfs.uni-tuebingen.de/resources/stts-1999.pdf, 1999 [accessed: 06.12.2019].

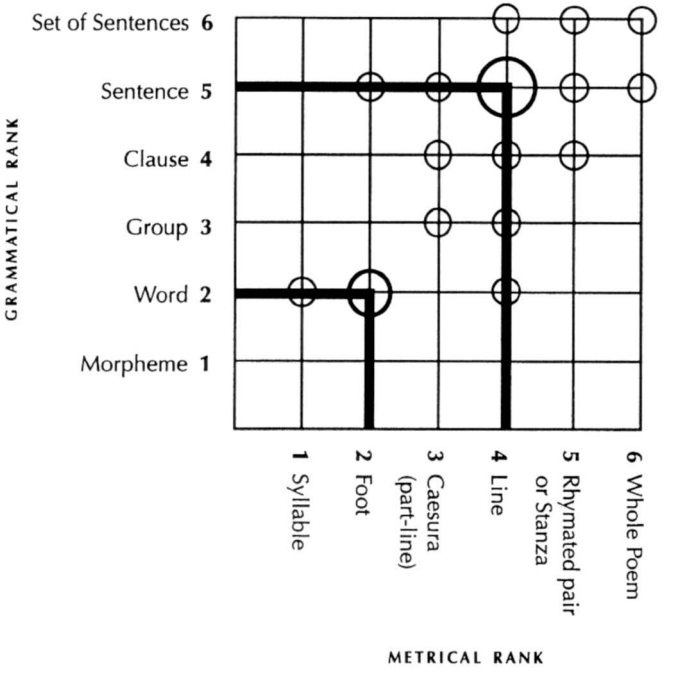
Method

This section describes the theoretical approaches used in the analysis and the computational approaches used for classification with the help of a web-based interface. It also describes the human-in-the-loop approach for classification and analysis of the corpus by using that interface.

Literary Analysis: Grammetrical Ranking and Rhythmic Phrasing

On the literary side, the classification is based on two theoretical approaches existing within the research field of free verse prosody: The 'grammetrical ranking'[13], and the 'rhythmic phrasing'[14].

Fig. 2: The vertical axis is the grammatical rank; the horizontal axis is the metrical rank. Intersection points help to identify the poem, for instance, the line arrangement (Wesling 1996).



13 Donald Wesling: *The Scissors of Meter: Grammetrics and Reading* (Ann Arbor, Michigan: University of Michigan Press, 1996).

14 Richard D. Cureton: *Rhythmic Phrasing in English Verse* (London: Longman, 1992).

The idea of grammetrical ranking has been developed by Donald Wesling, based on the key hypothesis that in poetry as a kind of versified language, the grammatical units (sentence, clause, group, word, morpheme) and the metrical units (syllable, foot, part-line, line, rhymated pair, stanza, whole poem) interact in a way for which Wesling finds 'scissoring' an apt metaphor (see Figure 1).

The idea of rhythmic phrasing was developed by Richard Cureton; it was based on Lerdahl and Jackendoff's 'Generative Theory of Tonal Music'.[15] Cureton divided the poetic rhythm into three components: metre, grouping and prolongation. Metre is about the perception of beats in regular patterns and grouping refers to the linguistic units gathered around a single climax or peak of prominence, quite similar to Wesling's ranking. Basically, Cureton's new idea is that of prolongation which refers to the anticipation and overshooting of a goal, the experience of anticipation and arrival.

In his groundbreaking study on 'The Prosodies of Free Verse' in 1971,[16] Donald Wesling distinguished five different prosodic types in the history of free verse prosody:

1. Whitmanic, referring to Whitman's adaptation of 'the biblical verset and syntax' in 'end-stopped lines .., with boundaries so often equivalent to those of larger units of grammar,' which Wesling sees as 'constitut[ing] the precomposition or matrix of free verse in English';
2. 'line-sentences,' as developed by Pound in Cathay on the basis of Ernest Fenollosa's theories of the sentence derived from study of Chinese;
3. dismemberment of the line, whereby the line becomes 'ground to the figures of its smaller units', and, as a sub-category, spatial dismemberment of the line by indentation, as by Williams in his triadic line verse;
4. systematic enjambment, whereby '[t]he lines.., are figures on the ground of the larger unit, the stanza'; and
5. dismemberment together with enjambment of the line, such that 'the middle units on the rank scale engage in a protean series of identity shifts as between figure and ground.'

Based on this typology extended by Wesling in later works,[17] we developed a fluency/disfluency spectrum based on nine different patterns. We illustrate this prosodic spectrum of fluency/disfluency by ranking these nine different poetic

15 Fred Lerdahl, Ray Jackendoff: *A Generative Theory of Tonal Music* (Cambridge, Mass: MIT Press, 1983).

16 Donald Wesling: "The Prosodies of Free Verse", in R. A. Brower (Ed.), *Twentieth-Century Literature in Retrospect* (Cambridge, Mass.: Harvard University Press, 1971).

17 Wesling, *Scissors of Meter*, 1996.

styles within the free verse spectrum. **(1) Cadence**: the cadence is the most fluent one. The basic idea of the cadence is the 'breath-controlled line' as an isochronous principle. Ezra Pound, who invented the idea of the cadence, was influenced by Chinese poetry that lacked any enjambments.[18] This explains the so-called line-sentence as the fundamental principle of the cadence. As different from the first class, more disfluent poems use 'weak enjambments' separating the nominal phrase and the verbal phrase of a sentence. Such 'weak enjambments' can be divided furthermore according to the emphases of the enjambments. **(2) Parlando**: the poems in the parlando style use 'weak enjambments' that do not emphasize the enjambments. **(3) Variable foot**: in contrast to parlando, the poems in the variable foot class emphasize the enjambments. These two classes are rather the fluent ones compared to those poems using 'strong enjambments', because the break in the reading is not really irritating as long as it is based on regular pauses in speech. Figure 1 shows an example of the rhythmical pattern 'variable foot' in the first two lines of a poem by the German poet Ernst Jandl: 'beschreibung eines gedichtes' (English: Description of a Poem). As it is obvious from the intensity or power contour, the poet makes a short stop after each colon in the sentence, imitating the regular pauses in speech by taking into account his breathing as a rhythmical feature (pause between words 'lippen' and 'ohne'). We used a text-speech aligner for the detection of word and sentence boundaries.

A more disfluent kind of poetry uses strong enjambments, which means it separates articles or adjectives from their nouns or even splits a word across lines, as in Paul Celan's poems. Poems using 'strong enjambments' can also be divided according to the emphases of the enjambments. **(4) Hook style** (German: Hakenstil): the poems in the hook style use 'strong enjambments' that do not emphasize the enjambments. **(5) Gestic rhythm**: in comparison to the hook style, the poems in the gestic rhythm emphasize the enjambments, meaning that the author makes an irritating break after each line when reading his poem, although the sentence should continue. This fifth pattern was invented by Bertolt Brecht and used in his later works. It had a huge impact on poets from the former German Democratic Republic.

Moving towards the radical disfluent pole, the next pattern is the ellipsis. **(6) Ellipsis**: it is the omission of one or more grammatically necessary phrases. This rhetorical figure can also affect the prosody of a poem, which has been observed in poems of Paul Celan. **(7) Permutation**: it is a conversion or exchange of words or parts of sentences or a progressive combination and rearrangement of linguistic-semantic elements of a poem, a principle that was very popular in German 'concrete poetry'.[19] Even more radical kinds of poetic disfluency have been developed in mod-

18 George Steiner: *After Babel—Aspects of Language and Translation* (Oxford: Oxford University Press, 1975), 358.

19 Ulrich Ernst: „Permutation als Prinzip in der Lyrik". *Poetica*, vol. 24, no. 3/4, 225–269, 1992.

ern 'sound poetry' by dadaistic poets like Hugo Ball and Schwitters or concrete poets like Ernst Jandl. Within the genre of sound poetry, there are two main patterns: **(8) Syllabic decomposition**, dividing the words into syllables; and **(9) Lettristic decomposition**, the last and most disfluent pattern, which divides the words into single characters, as for example in Ernst Jandl's famous poem 'schtzngrmm'.

Computational Analysis

We devised our computational methods with the primary concerns on *interpretability* (i.e., the ability of a method to explain what aspects of the data determine its behaviour) and the ability to deal with *data sparsity* (i.e., the limitation that there is too little data to fully train advanced machine learning algorithms).

Poetic Prosody Classification

Our project setup allows *supervised learning* where the parameters of a computational model are tuned to data (in our case poems) that have already been classified into a range of classes. The goal of the classifier is to extract those aspects of a poem that characterize it as part of a class, while neglecting other aspects, so that the resulting characterizations generalize to previously unseen poems.

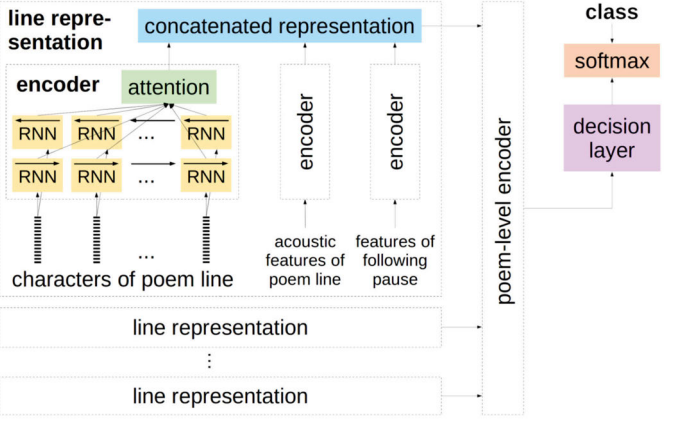
We first focused our efforts on building interpretable classifiers, such as decision trees, and on specific interpretable features, such as the presence or absence of a verb in a line.[20] Obviously, most poems consist of many lines and, therefore, the question of how to integrate knowledge across lines arises (e.g., what should be the proportion of lines that exhibit a certain feature to be deemed relevant?). Of course, there is no clear answer and the question of what aggregation of what features in a line yields the best performance assumes some significance (add to it aspects such as *speaking style* which cannot easily be measured by a discrete feature). We therefore focused more and more on deep-learning-based methods that help to discover latent or hidden structure in data across numerous dimensions (such as in the text, the speech, and other aspects of recitation). Particularly, based on the literary theory, we determined that pauses between lines were a critical aspect in addition to the speech and the text.

Deep learning is known to require a great amount of data and although the number of poems manually analysed in the course of the project may appear large, com-

20 Hussein Hussein, Burkhard Meyer-Sickendiek, Timo Baumann: "Tonality in Language: The 'Generative Theory of Tonal Music' as a Framework for Prosodic Analysis of Poetry", in: *Proceedings of the 6th International Symposium on Tonal Aspects of Languages (TAL) 2018*, Berlin, Germany, June 2018a; Hussein Hussein, Burkhard Meyer-Sickendiek Timo Baumann, "Automatic Detection of Enjambment in German Readout Poetry", in: *Proceedings of Speech Prosody*. Poznán, Poland, 2018b.

plex machine-learning models would require many orders of magnitude more data to be able to work at their full potential.

Fig. 3: Full model for deep-learning-based poetry style detection: each line is encoded character-by-character by a recurrent neural network (using GRU cells) with attention. Acoustic features of each line and the pause following up to the next line are encoded similarly. Per-line representations are concatenated and passed to a poem-level encoder. The final decision layer optimizes for the poem's class.



Also, the fact that the prosodic structure of a poem is reflected in many (or most) lines of a poem that belongs to a certain pattern, can be exploited by *hierarchical attention models*[21] which presuppose structure in a given document.[22] In our case, we know that a poem consists of lines and given text-speech alignment, we also know both the text and the speech audio that correspond to the line. Furthermore, given that the pause at the end of a line is stylistically relevant, for example, to differentiate different types of enjambment, we also encode the pause that separates a line from the next. We combine the information from these three aspects of a line (text, speech and following pause) using three recurrent neural networks (bidirectional GRUs) that furthermore use inner-attention (also called self-attention) so as to focus on what is most relevant in the line. In order to combine the information from all

21 Zichao Yang et al.: "Hierarchical attention networks for document classification", in: *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics*: Human Language Technologies. pp. 1480–1489, Association for Computational Linguistics, 2016.

22 In this case, "document" refers to a poem.

lines into our class assignment, we use yet another recurrent layer across the lines. This architecture is depicted in Figure 3. As a result, our classifier is able to single out the information that differentiates stylistic patterns as it is trained to be optimal towards differentiating the prosodic classes in the training material.

Given the fact that most of the lines exhibit the structural properties that yield a poem's prosodic classification, we use a two-stage training procedure, in which we first train the line-by-line classification in isolation and only afterwards do we address the full network including the final decision layer. In this way, we make most efficient use of our data and successfully tackle the issue of data sparsity. We furthermore improve results by pretraining part of our representation encoders using poetic texts (from www.deutschestextarchiv.de) and general speech audio. We set aside testing material and find that our classifiers achieve high performance.

After establishing that deep learning could be used to classify lines of poems according to poetic prosodic style,[23] we extended it to experiments on classifying poetry by using a hierarchical attention network.[24] We found that free verse prosodies could be classified along a fluency continuum and that the classifier's wrong classifications cluster along this continuum. Through ablation studies (i.e., leaving out certain features to find their relative importance for the final results), we found that, depending on the classes analysed, the acoustic realization of the speech itself is highly relevant for classification, so is the realization of the pause following each line of the poem. In fact, our classifier is highly reliable for determining different kinds of enjambments and an additional annotation of enjambments is unnecessary in the poems (unlike what we had originally hypothesized). We also tried traditional classification approaches by using engineered features and, in a comparison with our deep-learning-based method, found that they did not work as well although they had direct access on theoretically grounded features.[25] We also found that features of a poetic line (such as enjambment) could be identified by a neural network and that their annotation did not improve the overall poem classification.[26] Finally, we integrate engineered features into the deep-learning approach which further boosts

23 Burkhard Meyer-Sickendiek, Hussein Hussein, Timo Baumann: "Recognizing Modern Sound Poetry with LSTM Networks", in: *Proceedings of Elektronische Sprachsignalverarbeitung (ESSV)*. TUDpress, 192–199, 2018.

24 Timo Baumann, Hussein Hussein, Burkhard Meyer-Sickendiek: "Style Detection for Free Verse Poetry from Text and Speech", in: *Proceedings of COLING*. Santa Fe, USA, 1929–1940, 2018a.

25 Timo Baumann, Hussein Hussein, Burkhard Meyer-Sickendiek: "Analysis of Rhythmic Phrasing: Feature Engineering vs. Representation Learning for Classifying Readout Poetry", in: *Proceedings of the Joint LaTeCH&CLfL Workshop*. Santa Fe, USA, 44–49, 2018b.

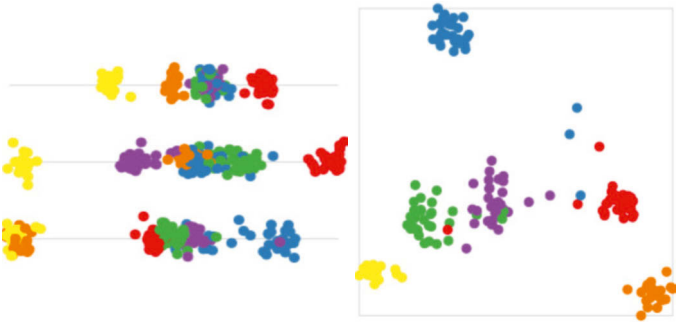
26 Timo Baumann, Hussein Hussein, Burkhard Meyer-Sickendiek: "Analysing the Focus of a Hierarchical Attention Network: The Importance of Enjambments When Classifying Postmodern Poetry", in: *Proceedings of Interspeech*. Hyderabad, India, 2162–2166, 2018c.

performance.[27] We refer the interested reader to our individual publications for details on the methods used.

Visualization of Results

Our results can provide multiple kinds of visualization which may help literary analysis. Visualization is particularly useful for deep-learning-based approaches as it provides some understanding of the inner workings of the model that is otherwise hard to understand. For example, representations of a poem (to be used for classification) can be forced into lower dimensions and the placement of poems in these dimensions can be visualized in 1D, 2D or 3D[28] spaces, as some are shown in Figure 4. The analysis of outliers (e.g., the one red poem among many green poems) has proved to be useful for literary analysis.

Fig. 4: Visualization of low-dimensional mappings of the poems contained in six classes (Left: Results depend on random initialization and three results are shown).

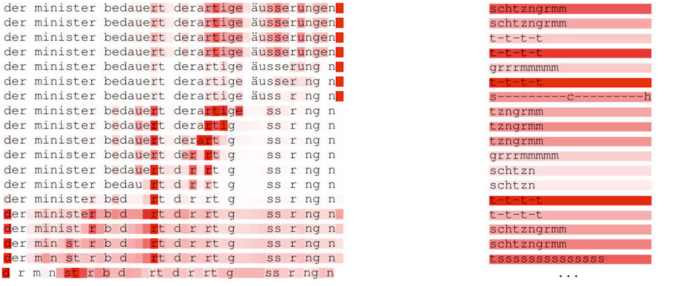


The attention mechanism in our model determines what the model pays attention to in a certain poem and this can be exploited for visualization. However, it needs to be noted that the attention is primarily focused on those aspects that allow the model to differentiate classes, not to describe classes. It can nonetheless shed light on classifier decisions, as can be seen in Figure 5.

27 Hussein Hussein, Burkhard Meyer-Sickendiek, Timo Baumann: "Free Verse and Beyond: How to Classify Post-modern Spoken Poetry", in: *Proceedings of Speech Prosody*. Tokyo, Japan, 690–694, 2020.

28 An interactive version of a 3D plot of poems analysed in this way is available at https://web.archive.org/web/20210404220806/https://timobaumann.bitbucket.io/colingfreeversepoetry/.

Fig. 5: Visualization of attention in two lettristic poems: (a) attention to characters within the line, (b) attention to lines (including the audio) in the poem.



(a) Jaap Blonk: Der Minister I (b) Ernst Jandl: schtzngrmm

Finally, a different form of visualization concerns the presentation of our subject matter—the free verse prosody—by using data from our partner *Lyrikline*. Meanwhile *Lyrikline* has visualized the classes we investigated on the website of *Lyrikline*, so an own link to *Lyrikline* represents our classes. This form of visualization is of course so important that our results can now be used for teaching poetics and prosody.

Iterative Human-in-the-Loop Approach

Above, we have described the methods we used for literary and computational analysis, respectively. We have selected poems from *Lyrikline* that form one of the multiple patterns and we have built computational models that are able to classify poems based on these patterns with relatively high performance. However, we have already mentioned the issue of data sparsity and it is tedious and prone to oversight errors in manually selecting poems that belong to a pattern. We therefore come back to our original ideas[29] and combine both parts of the project to iteratively analyse many more poems in the *Lyrikline* collection. We have implemented an interface for *humanist-in-the-loop* classification and analysis[30] that allows a literary scholar to manually assign poems that have been automatically classified to a certain class. We have gradually applied this method to cover the full corpus.

29 Timo Baumann, Burkhard Meyer-Sickendiek: "Large-scale Analysis of Spoken Free-verse Poetry", in: *Proceedings of Language Technology Resources and Tools for Digital Humanities (LT4DH)*, 125–130, Osaka, Japan, December 2016.

30 Timo Baumann et al.: "A Tool for Human-in-the-Loop Analysis and Exploration of (not only) Prosodic Classifications for Post-modern Poetry", in: *Proceedings of INF-DH*. Kassel, Germany. Gesellschaft für Informatik, 151–156, 2019.

The main contribution is the integration of a human-in-the-loop process via a web-based interface. The web interface uses the results of the automatic classifier and makes as much information about the decision making process available as is possible. In so doing, the interface also includes the model's confidence (the trust of the model in its judgment) and its attention (places in the text and speech that the model deems particularly relevant for the judgment). The interface enables a philologist to explore the classification results and potentially correct or enhance them.

During the analysis of our data we repeatedly find a pattern that is not yet defined and we include it in our corpus. This pattern must again go through the quantification process. If there are more than 25 examples of this new pattern, then it is a relevant pattern of modern and postmodern free verse prosody. That is, we assume a threshold value that helps us to compensate for the difference between qualitative and quantitative analysis.

Our interface for poem classification allows us to add new classes and change prior class assignments in the light of new classes. For example, it allows splitting a class into multiple sub-classes and then retraining the computational models as the exploration of the corpus proceeds in order to help with further analyses.

In the end and over the course of the project, especially in the literary studies part of the project, a total of 18 rhythmic patterns have been developed: Longline Poems; Parlando; Flows; Free Associations; Cadence; Unstressed Enjambments; Variable Foot; Sprungrhythm; Syncopations; Rubato; Permutation; Gestic Rhythm (= Stressed Enjambments); Cut-ups; Dialect Poems; Ellipses; Staccato; Syllabic Decomposition; and Lettristic Decomposition. All these patterns are explained and illustrated by a number of examples on the *Lyrikline* website. Also, a monograph furthermore explains the theoretical framework, the patterns mentioned and the development of free verse prosody in modern and postmodern poetry.[31]

Conclusion

To conclude, we first summarize our approach and discuss the outcome of our project before we briefly describe some future work.

In the Rhythmicalizer project we have melded computational and literary methods to introduce ideas of free verse poetry into the debate on present-day post-modern German poetry. We have linked the ideas of grammetrical ranking as well as rhythmic phrasing and found a fluency/disfluency continuum along which many poems can be classified.

31 Burkhard Meyer-Sickendiek: *Hörlyrik. Eine interaktive Gattungstheorie* (München/Paderborn: Brill/Fink, 2020).

In the computational part of the project, we have shown that it is possible to build classifiers based on speech and text that are able to differentiate the prosodic patterns despite the high complexity and sparsity of the given poems. The resulting classifiers show patterns that underline the fluency/disfluency continuum.

We have finally built a *humanist-in-the-loop* interface that enables us to further analyse and classify a large portion of the poems available on *Lyrikline by* using a data-driven methodology to work (read and listen) through thousands of poems.

In our project, we were able to analyse a large portion of the German poems on *Lyrikline* and to trace the development of free-verse prosody in modern and post-modern poetry.

A central limitation of our approach is the reliance on the differentiation of classes. There are always uncertainties in classification, in particular for something that is as stylistically diverse and continuous as poetry. They primarily consist in the polyvalence of poems which can never really be assigned to a single class or characteristic. On the contrary, in good poems, several patterns or characteristics can often be found, which is why the assignment to exactly one pattern is always fraught with uncertainties. In this case, it is important to add considerable weight to the most important characteristics from a literary historical point of view. To give an example: In expressionist poems, the so-called 'Reihenstil' (=sequence style) is the most important feature, that is, the sequence of events that are heterogeneous in themselves. There are, however, other characteristics that characterize an expressionist poem. This issue could have been overcome by introducing a hierarchy of classes or by establishing a multi-class labelling approach. While potentially feasible, such approaches complicate the computational analysis and only partially alleviate the problem. We have therefore limited the analysis to plain classification.

Any classification of a work of art is always an abstraction, neglect of details. We could only classify the poems on *Lyrikline* by following a strategy of simplification and reduction. This strategy became even more extreme when it came to the characteristics of computational learning, as these characteristics were necessarily simpler and more reduced than the philologist's observations. At this point the calibration or threshold value is again important: Is the detailed observation possibly not a single case but a recurring pattern?

Another limitation is the type of machine learning used for classification. A neural network (and most other classification algorithms) does not model the classes and their properties, but merely differentiates between classes. Therefore, any analysis of the learned classifiers is limited and differs significantly from a human perspective which aims at forming classes based on their inner coherence. The hermeneutical idea of learning, in contrast, is way more based on the idea that different 'horizons' of human experiences can be mingled during the learning phase. These 'horizons' obviously do not exist in machine learning. Here we refer to Hans Georg Gadamer's concept of 'merging horizons', which describes the process

of understanding as a mixture of classical texts and works of art with a contemporary horizon, that is, a process of understanding between different cultures. This form of learning is naturally missing in artificial intelligence.

The second central conflict concerns the accentuation of our classification by the provider *Lyrikline* or by the authors of *Lyrikline* classified by us. It is not always acceptable for lyricists or artists to see their work classified and assigned in a certain way. As a rule, this may only be done with artists or writers who died long ago. Living artists often refuse to be classified. We hope (and some positive examples underpin this) that our data-driven endeavour nonetheless convinces them that classification of their work does not diminish or demolish their work.

Over the course of the project, multi-modal processing has become more commonplace (and the tools that can freely be used are easier to use) in the Digital Humanities, and deep-learning-based approaches are also more common. We hope that our project sets a good example for the integration of multi-modal data with an iterative human-in-the-loop approach to data analysis that yields interesting problems and results both in computational studies and in the humanities.

At present, we are involved in transferring our research results to humanities teaching at the high school level. Our interactive web-based human-in-the-loop project can be used by students to (a) learn about present-day poetry and the various forms that it takes, to (b) reflect about language use in poetry, for example, to learn how to differentiate between enjambments based on their use in a poem by singling out and paying attention to details, and to (c) understand how a topic as liquid and hard to grasp as poetry can nonetheless be differentiated, categorized, classified or hierarchized. These competences go far beyond their immediate use in poetry analysis and are highly relevant for future humanities scholars in a computational age.

Bibliography

Baumann, Timo, Arne Köhn, and Felix Hennig: "The Spoken Wikipedia Corpus Collection: Harvesting, alignment and an application to hyperlistening", *Language Resources and Evaluation*, vol. 52, no. 2, 303–329, 2018, special Issue representing significant contributions of LREC 2016.

Baumann, Timo, Hussein Hussein, and Burkhard Meyer-Sickendiek: "Style Detection for Free Verse Poetry from Text and Speech", in: *Proceedings of COLING*. Santa Fe, USA, 1929–1940, 2018a.

Baumann, Timo, Hussein Hussein, and Burkhard Meyer-Sickendiek: "Analysis of Rhythmic Phrasing: Feature Engineering vs. Representation Learning for Classifying Readout Poetry", in: *Proceedings of the Joint LaTeCH&CLfL Workshop*. Santa Fe, USA, 44–49, 2018b.

Baumann, Timo, Hussein Hussein, and Burkhard Meyer-Sickendiek: "Analysing the Focus of a Hierarchical Attention Network: The Importance of Enjambments When Classifying Post-modern Poetry", in: *Proceedings of Interspeech*. Hyderabad, India, 2162–2166, 2018c.

Baumann, Timo, Hussein Hussein, Burkhard Meyer-Sickendiek, and Jasper Elbeshausen: "A Tool for Human-in-the-Loop Analysis and Exploration of (not only) Prosodic Classifications for Post-modern Poetry", in: *Proceedings of INF-DH*. Kassel, Germany. Gesellschaft für Informatik, 151–156, 2019.

Baumann, Timo and Burkhard Meyer-Sickendiek: "Large-scale Analysis of Spoken Free-verse Poetry", in: *Proceedings of Language Technology Resources and Tools for Digital Humanities (LT4DH)*, 125–130, Osaka, Japan, December 2016.

Cho, Kyunghyun, Bart van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio: "Learning phrase representations using rnn encoder–decoder for statistical machine translation", in: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1724–1734, Doha, Qatar. Association for Computational Linguistics, 2014.

Cureton, Richard D.: *Rhythmic Phrasing in English Verse*. Longman, 1992.

Ernst, Ulrich.: "Permutation als Prinzip in der Lyrik". *Poetica*, vol. 24, no. 3/4, 225–269, 1992.

Geitner, Ursula: *Die Sprache der Verstellung. Studien zum rhetorischen und anthropologischen Wissen im 17. und 18. Jahrhundert*, De Gruyter, 1992.

Hartman, Charles O.: *Free Verse: An Essay on Prosody*. Princeton University Press, 1980.

Hussein, Hussein., Burkhard Meyer-Sickendiek, and Timo Baumann: "Tonality in Language: The "Generative Theory of Tonal Music" as a Framework for Prosodic Analysis of Poetry", in: *Proceedings of the 6th International Symposium on Tonal Aspects of Languages (TAL) 2018*, Berlin, Germany, June 2018a.

Hussein, Hussein, Burkhard Meyer-Sickendiek, and Timo Baumann, "Automatic Detection of Enjambment in German Readout Poetry", in: *Proceedings of Speech Prosody*. Poznán, Poland, 2018b.

Hussein Hussein, Burkhard Meyer-Sickendiek, and Timo Baumann: "Free Verse and Beyond: How to Classify Post-modern Spoken Poetry", in: *Proceedings of Speech Prosody*. Tokyo, Japan, 690–694, 2020.

Jandl, Ernst: *der gelbe hund*, Darmstadt und Neuwied: Luchterhand, 1985.

Katsamanis, Athanasios, Matthew Black, Panayiotis G. Georgiou, Louis Goldstein, and Shrikanth S. Narayanan, "Sail Align: Robust Long Speech-Text Alignment", in: *Proc. of Workshop on New Tools and Methods for Very-Large Scale Phonetics Research*, 2011.

Lerdahl, Fred and Ray Jackendoff: *A Generative Theory of Tonal Music*. Cambridge, Mass. MIT Press, 1983.

Meyer-Sickendiek, Burkhard.: *Hörlyrik. Eine interaktive Gattungstheorie*. Brill/Fink, Paderborn, München 2020.

Meyer-Sickendiek, Burkhard andHussein Hussein: "The New German Copyright Law for Science and Education (UrhWissG): Consequences for DH-Projects Working with Non-Academic Partners", in: *Proceedings of the 15th Conference on Knowledge Organization WissOrg'17 of the German Chapter of the International Society for Knowledge Organization (ISKO)* 2017, Berlin, December 2017.

Meyer-Sickendiek, Burkhard, Hussein Hussein, and Timo Baumann: "Recognizing Modern Sound Poetry with LSTM Networks", in: *Proceedings of Elektronische Sprachsignalverarbeitung (ESSV)*. TUDpress, 192–199, 2018.

Rafferty, Anne N. and Christopher. D. Manning, "Parsing Three German Treebanks: Lexicalized and Unlexicalized Baselines", in: *Proceedings of the Workshop on Parsing German*, Stroudsburg, PA, USA: Association for Computational Linguistics, 2008, 40–46.

Schiller, Anne, Simone Teufel, Christine Stöcker, and Christine Thielen, "Guidelines für das Tagging deutscher Textcorpora mit STTS (Kleines und größes Tagset)", Available on http://www.sfs.uni-tuebingen.de/resources/stts-1999.pdf, 1999, last accessed on 06. December 2019.

Sjölander, Kånder and Jonas Beskow, "Wavesurfer–an open source speech tool", in: *Sixth International Conference on Spoken Language Processing*, 2000.

Steiner, George: *After Babel—Aspects of Language and Translation*. Oxford University Press, 1975.

Walker, Willie, Paul Lamere, Philip. Kwok, Biksha Raj, Rita Singh, Evandro Gouvea, Peter Wolf, and Joe Woelfel, "Sphinx-4: A Flexible Open Source Framework for Speech Recognition", Mountain View, CA, USA, Tech. Rep., November 2004.

Wesling, Donald: "The Prosodies of Free Verse", in *Twentieth-Century Literature in Retrospect*. R. A. Brower (Ed.), Harvard University Press, 1971.

Wesling, Donald: *The Scissors of Meter: Grammetrics and Reading*. University of Michigan Press, 1996.

Yang, Zichao, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy: "Hierarchical attention networks for document classification", in: *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. pp. 1480–1489, Association for Computational Linguistics, 2016.

#MACHINE LEARNING

Like many other digital humanities terms, #Machine Learning appeared as neologism specifically in the field of computing during the post-war period. It focuses on "the capacity of computers to learn and adapt without following explicit instructions, by using algorithms and statistical models to analyse and infer from patterns in data." * The definitions by the mixed methods projects sketch broader and deeper fields of issues. They point to different processes of learning in the area of computing in comparison with human learning and trace the structure as well as process of the learning, thus connecting machine learning to #Human-in-the-loop.

First, they differentiate between supervised and unsupervised learning: "The supervised learning maps an input – based on example input-output pairs– to an output. In the unsupervised learning, there are no pre-existing labels and it tries to find previously unknown patterns in data set." Different strategies of #Machine Learning address different information or research interests when the "algorithms use manually engineered features or automatically learnt representation" (Rhythmicalizer). As such, the decision for a specific alignment of the algorithm forms a frame for the possibility and impossibility of results.

In addition to it, some projects underline that the application "of differentiable and parameterized mathematical functions" for "pattern recognition is effective at highly specific pre-formulated tasks of high labour intensity, but fails at transferring gained insights to different non-formulated contexts" (ANCI). Here, different roles of the human element emerge, ranging from the creation of the data and the algorithms to supervising, to potentially re-ordering, and finally, to interpreting the learning process.

* "machine, n.". in: Oxford English Dictionary (OED), Third Edition, March 2000; most recently modified version published online June 2021, https://www.oed.com/ [accessed 25.10.2022].

Title: anci – analysing networked climate images
Team: Birgit Schneider, Frank Heidmann, Thomas Nocke, Janna Kienbaum, Paul Heinicker
Corpus: We created our own research database from scratch via automated sampling of thousands of climate change images from the web. We filtered these images by certain keywords because we were interested in so called catch images, as specific recurring visual manifestations, like the Hockey Stick graph, the burning earth or protest photography.
Field of Study: Visual Studies, Media Studies, Human Cartography, Environmental Humanities, Interface Design
Institution: University of Potsdam, University of Applied Sciences Potsdam, Potsdam Institute for Climate Impact Research
Methods: The project searches for methods to conquer the sheer mass of climate imagery on the web to technically gather and process data. By quantitative means, we focus on approaches of visual analytics with the help of various visualisation techniques but also purely algorithmic approaches ranging from web scraping, computer vision, natural language processing and their extension via machine learning. These algorithmic processes were critically reflected and are in interaction with qualitative approaches to analyse the content. Hence, we used formal-stylistic or iconic, iconographic, iconological analysis methods as well as diagrammatic approaches. In addition, framing analysis was transferred from communication science to visual climate communication in order to analyse image content and its contextualisation for frames or specific statements and arguments.
Tools: There are various tools used in the project. Our toolset includes a web scraping application called *anci miner* – which we developed and programmed as main-outcome of our research endeavor, a visualisation interface for high-dimensional data sets with an additional cluster feature and a data visualisation interface which allows quick insights to main themes within a text corpus.
Technology: For the implementation of natural language processing, computer vision and machine learning algorithms we relied on open-source Python frameworks, such as OpenCV or Google's tensorflow. The resulting tools are custom-made web-based applications built with JavaScript and also enriched by open source add-ons

Interpreting Climate Images on the Internet: Mixing Algorithmic and Interpretive Views to Enable an Intercultural Comparison (ANCI)

Birgit Schneider, Thomas Nocke, Paul Heinicker, Janna Kienbaum

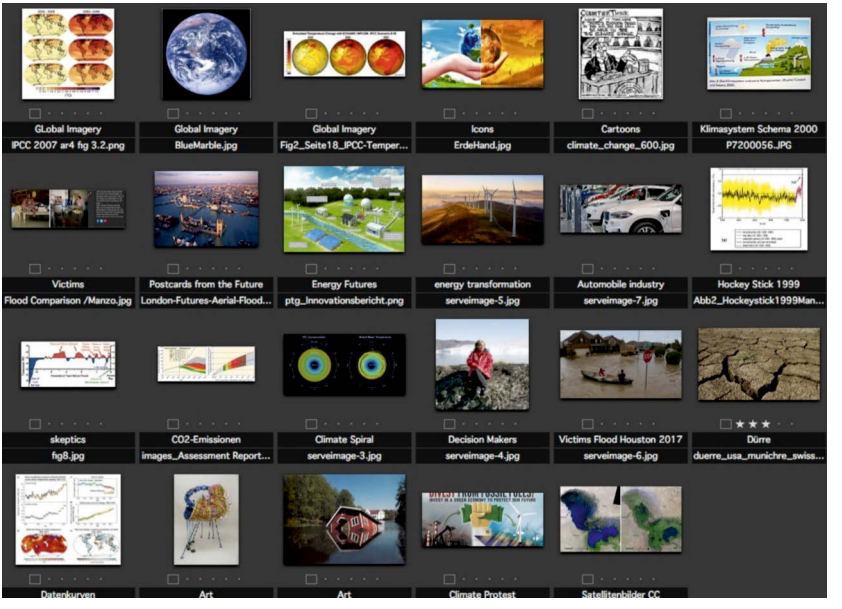
Abstract *Countless climate images are in circulation on the internet, such as burning globes, polar bears and photos of global climate impacts. These images are networked and generate a specific view of climate change. Our case study engages in an intercultural image comparison based on Google Image queries. As an interdisciplinary team of experts drawn from art history, media studies, interface design and computer graphics, our goal was to use a combination of qualitative and quantitative image analyses to explore the predominant discourses of digitised visual climate communication on the web. To this end, we automated the analysis of different formal features of climate images (such as colour values, density and composition) with the aid of computer-driven methods (such as computer vision and machine learning) to build a corpus of thousands of images. Our focus was on image similarities, a concept shared by both image theory and computer analysis. In this chapter, we elucidate the outcome of our research on a conceptual and technical basis. The core issue addressed here is the manner in which art-historical methods (such as iconography and the concept of visual framing) are transformed when using computer-generated methods of computer vision and machine learning to analyse image similarities. This chapter focuses on our various insights while also reflecting on the general question of networked images on a methodological level. Ultimately, we were able to identify the promising potential but also the key limits of algorithmic image recognition and sorting when using machine learning to study images on the internet.*

Introduction

The research project *anci*—an acronym for *a*nalysing *n*etworked *c*limate *i*mages—addresses the topic of climate change and its visual impact in communication. Ever since the popularisation of the World Wide Web, climate images have increasingly circulated online. Some motifs became deeply rooted in cultural web memory and now influence the world view of global warming in the realm between politics, science, art and popular culture. For the purpose of our research, we compared and

analysed these images and search for similarities with the help of 'catch images' (analogous to catch words), a term coined by the cultural scholar Aby Warburg (1866–1929).[1] The term refers to images that dominate the current global climate picture language in scientific, special or political contexts, such as photographs of the globe as a 'Blue Marble' and temperature curves of global warming in a hockey stick shape (Fig. 1).

Fig. 1: Selection of catch images. Excerpt from the Expression Media Pro database 'Klimathek' by Birgit Schneider. Images in this database are sorted by means of hand-tagged keywords.



By using a method of arranging image reproductions on large panels in order to understand their relatedness, Warburg sketched out a key method for comprehending networked images in the early twentieth century.[2] Today, in the age of image databases, his approach provides a great model for digital image research. While

1 "The interest here is primarily in the transfer of the image into the sphere of everyday political life as well as the iconography of these everyday images, which are apostrophized as public images or—to use a fitting word from Aby Warburg—as striking images." Diers, *Schlagbilder: Zur politischen Ikonographie der Gegenwart*, 7 / Diers, *Catchimages: On the Political Iconography of the Present*, 7 (translated using DeepL).

2 See Aby Warburg, *Der Bilderatlas Mnemosyne* / The image atlas Mnemosyne (translated using DeepL) (Berlin: Akademie Verlag, 2008).

analysing networked images, similarity becomes systematically and methodically important as a concept, because it serves as a way to group images. Already in the late nineteenth century, the systematic comparison of images shaped a medial dispositive in academic art instruction.[3] Today, it is increasingly being used as an "intellectual operation"[4] in computer-controlled calculation procedures, such as computer vision and machine learning. We used Warburg's methodology as a starting point for our digital image studies.

Climate images sorted by Google Images

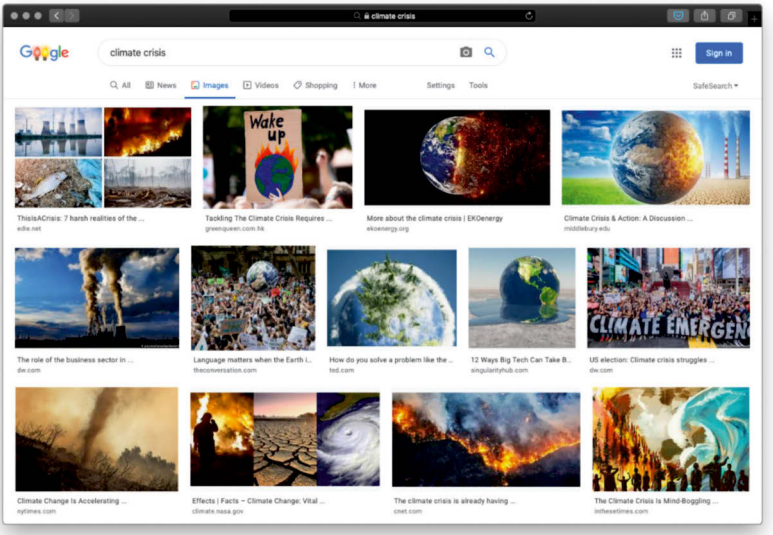
We applied our mixed-methods approach to the general question of climate communication, as our interest was limited to mainstream image communication in cultural spaces. We therefore deliberately did not examine social media; instead, we were interested in the largest possible, extensively used and highly formalised media spaces.

In contrast to other cultural data presented in this book, climate images on the internet are 'native' digital sources that are interconnected via HTTP protocol and have metadata. If you use Google to search for climate images, you will encounter a fairly standardised visual language for representing climate change—one that is dominated by globes, smoking chimneys and impacted landscapes (Fig. 2). At the same time, if you look at larger amounts of searched images, there are considerable differences in the preferred types of images (e.g., photos, maps, curves, pictograms or cartoons) and also in the framing of climate change.

3 See Heinrich Dilly, "Lichtbildprojektion—Prothese der Kunstbetrachtung" / "Photo projection—prosthesis of art perception" (translated using DeepL), in: Horst Bredekamp et al. (ed.): *Kunstwissenschaftliche Untersuchungen des Ulmer Vereins, Verband für Kunst- und Kulturwissenschaften* (Gießen: Anabas, 1975), 153–172.

4 Felix Thürlemann, "Bild gegen Bild: Für eine Theorie des Vergleichenden Sehens" / "Image against Image: For a Theory of Comparative Seeing" (translated using DeepL), in: Aleida Assmann, Ulrich Gaier, Gisela Trommsdorff (eds.): *Zwischen Literatur und Anthropologie—Diskurse, Medien, Performanzen* (Tübingen: Narr, 2005), 176.

Fig. 2: Research query via Google Image Search using the keyword 'climate crisis' taken on 2nd December 2020, Berlin, Germany, screenshot.



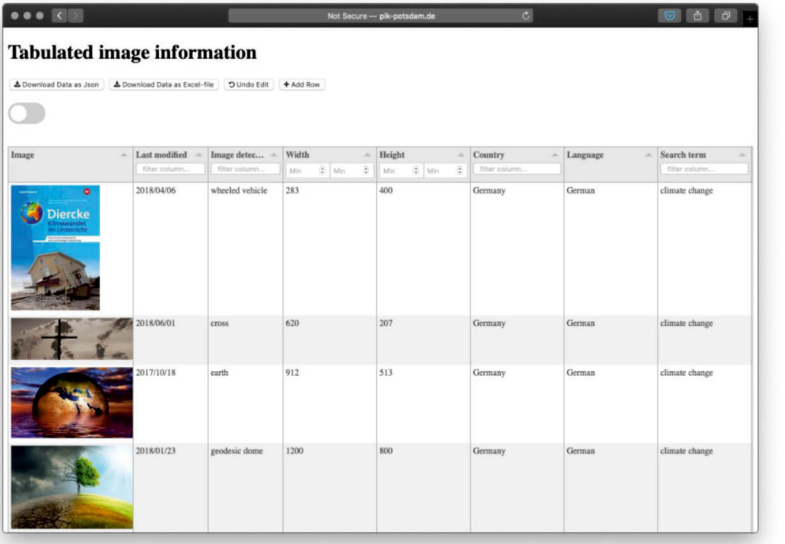
Since the 1990s, the automated indexing of websites has provided some of the main interfaces with which we access the World Wide Web. Alphabet Inc.'s Google is considered to be the dominant search engine in many parts of the world; it is also the most visited website ever, with several billion searches being clicked every day.[5] Despite acknowledging that Google's web service does not provide a neutral representation of hierarchies on the World Wide Web, we used the image query of Google Images as the basis of our image corpus. With the aid of the *Tor* browser and its VPN tunnelling capabilities, we technically imitated search queries from different cultural regions relating to climate issues. In total, we extracted roughly 16,000 images using eight keywords (e.g., climate change, climate change disaster, global warming) throughout seven cultural regions.[6] These images were accessed and edited via an online interface based on the Tabulator framework (Fig. 3).[7]

5 Statista, "Number of explicit core search queries powered by search engines in the United States as of April 2021".

6 Australia, USA, Brazil, Germany, Kenya, United Arab Emirates, Bangladesh. Instead of countries, we speak of locales in this regard. These locales serve as a cultural spatial unit. A locale combines the idea of a specific language with a region or country.

7 https://tabulator.info/.

Fig. 3: Web interface of our database of extracted images from Google Images according to specific keywords. Screenshot, 2020.



Methodology

In our research design, we applied specific combinations of methodological approaches. We oriented ourselves towards the mixed-methods designs put forward by the empirical social scientist John Creswell.[8] Two central meta-methods were of crucial importance in our research project: on the one hand, the method of *image comparison and image similarity* (qualitative) and, on the other, the method of *data visualisation*.

For our qualitative analysis of images, we combined different methods and disciplinary perspectives. On the level of *Bildwissenschaft* (the term for this German field

8 We oriented ourselves to Sequential Exploratory Design, the Sequential Explanatory Design and the Embedded or Nested Design. Cf. Creswell; Clark, *Designing and Conducting Mixed Methods Research*. For our image study on cross-cultural image comparison with Google, we decided to use Creswell's 'embedded design'. Although quantitative and qualitative approaches are still strictly separated here, these can be considered to be more in dialogue due to the simultaneity and dependence of the design. The quantitative perspective dominates the qualitative perspective, although interventions within the ongoing methodological approach can make it more flexible.

of image research can be translated as visual or image studies) and art history, we used image comparison in iconographic and iconological analysis methods,[9] but we also drew on rhetoric-semiotic[10] approaches. For the analysis of the distribution and migration of images, we turned to media-science approaches and collective image discourses.[11] For our cross-cultural image comparison with Google Images, we drew on the framing method[12] as used in communication science in tandem with the art-scientific theory of iconology as put forward by Erwin Panofsky.[13]

Frames generate their power through a symbolic charge. Images, in particular, make it possible to quickly and easily assimilate content that—in comparison to text—tends to lead latently to specific constructions of meaning: "Images are powerful framing tools because they are less intrusive than words and as such require less cognitive load. Therefore, peripheral rather than central processing may be activated and audiences may be more likely to accept the visual frame without question."[14] The visual frames can be derived from the image contents in combination

9 See Panofsky, *Sinn und Deutung in der bildenden Kunst / Meaning and interpretation in the visual arts* (translated using DeepL) (Köln: Dumont, 1978 [1955]).

10 See Lynda Walsh, "The Visual Rhetoric of Climate Change", *in: Wires Climate Change*, Volume 6, Issue 4, 361–368, 2015; See Lynda Walsh, "Tricks, Hockey Sticks, and the Myth of Natural Inscription: How the Visual Rhetoric of Climategate Conflated Climate with Character", in: Thomas Nocke, Birgit Schneider (eds.): *Image Politics of Climate Change Visualizations, Imaginations, Documentations* (Bielefeld: Transcript, 2014), 81–105; See also Birgit Schneider, *Klimabilder: Eine Genealogie globaler Bildpolitiken von Klima und Klimawandel / Climate Images: A Genealogy of Global Image Politics of Climate and Climate Change* (translated using DeepL) (Berlin: Matthes & Seitz, 2018).

11 See Jürgen Link, "Literaturanalyse als Interdiskursanalyse: Am Beispiel des Ursprungs literarischer Symbolik in der Kollektivsymbolik" / "Analysis as Interdiscourse Analysis: Using the Example of the Origin of Literary Symbolism in Collective Symbolism" (translated using DeepL). In: Jürgen Fohrmann, Harro Müller (eds.): *Methoden diskursanalytischer Ansätze* (Frankfurt a.M.: Suhrkamp, 1988).

12 See Robert Entman, "Framing: Toward Clarification of a Fractured Paradigm", *in:* Mark R. Levy (ed.): *Journal of Communication* (Volume 43, Issue 4, 1993); Jörg Matthes, *Framing* (Baden-Baden: Nomos, 2014); Elke Grittman, "Visual Frames—Framing Visuals. Zum Zusammenhang von Diskurs, Frame und Bild in den Medien am Beispiel des Klimawandeldiskurses" / "Visual Frames – Framing Visuals. On the Connection between Discourse, Frame and Image in the Media Using the Example of Climate Change Discourse" (translated using DeepL), in: Stephanie Geise, Katharina Lobinger (eds.): *Visual Framing: Perspektiven und Herausforderungen der visuellen Kommunikationsforschung* (Köln: Halem, 2015), 95–116; Stephanie Geise, Katharina Lobinger, *Visual Framing: Perspektiven und Herausforderungen der Visuellen Kommunikationsforschung / Visual Framing: Perspectives and Challenges in Visual Communication Research* (translated using DeepL) (Köln: Halem, 2015); Daniela V. Dimitrova, Lulu Rodriguez, "The levels of visual framing", *in:* Journal of Visual Literacy, (30:1), DOI: 10.1080/23796529.2011.11674684, 48–65.

13 See Panofsky, *Sinn und Deutung in der bildenden Kunst / Meaning and interpretation in the visual arts* (translated using DeepL).

14 Dimitrova; Rodriguez, The levels of visual framing, 50.

with their mode of representation (e.g., shooting angle of a photograph): "Visuals, like text, can operate as framing devices insofar as they make use of various rhetorical tools—metaphors, depictions, symbols—that purport to capture the essence of an issue or event graphically."[15]

We chose the framing approach so as to systematically do justice to the vast amount of Google images. Here we combined deductive and inductive approaches. The framing categories used in communication studies and developed by communication scholar Robert Entman allowed us to deductively filter and analyse climate images by means of algorithmic image analysis, that is, by emphasising a 'problem', 'cause', 'solution' or 'moral aspect'. Also, we were able to inductively derive frames from the image corpus based on the algorithmically generated image clusters.

Another method which helped us analyse algorithmically generated image clusters was *data visualisation*. We understand data visualisation as a mediator between computer science and art history or *Bildwissenschaft* (image studies). It makes the inherent complexity of data and abstract processes visible and thus tangible for the different perspectives and content analysis. Indeed, data visualisation methods were key to our analysis on many levels. They helped us when visualising statistical results using diagrams, such as tables or interfaces, but also when clustering corpora as image plots (structured overviews by different features). And, finally, they made it possible for us to illustrate the processes of machine learning itself. The resulting data visualisation generated by our research combined all previous algorithms into one image.

The entire process—beginning with image data collection (web scraping) and ending with visualisation—could be portrayed in the form of a pipeline (Fig. 4). In the first step, we compared the extracted images for similarities by means of a pretrained machine-learning model. The result was a high-dimensional data set that could not be understood at first sight, because it was too complex. For this reason, we reduced it to a dimension level that was perceptible to humans by using another machine-learning algorithm called *t-SNE*.[16] We then visualised the reduced data set and grouped it using a clustering algorithm called *k-means*.[17] The potential offered by the visualisation comes in the form of an interface between humans and mechanical processes, thus providing crucial access for all the described processes (Fig. 5). The challenge lies in working with such highly abstracted images that can certainly mask the underlying processes to some extent.

15 Dimitrova; Rodriguez, The levels of visual framing, 51.

16 See Geoffrey Hinton and Laurens van der Maaten, Visualizing Data using t-SNE (2008).

17 See J. MacQueen, "Some methods for classification and analysis of multivariate observations", in: *Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability*, (1, 1967), 281–297.

Fig. 4: The tools and algorithms that were used and the data that were visualised. Image: anci 2019.

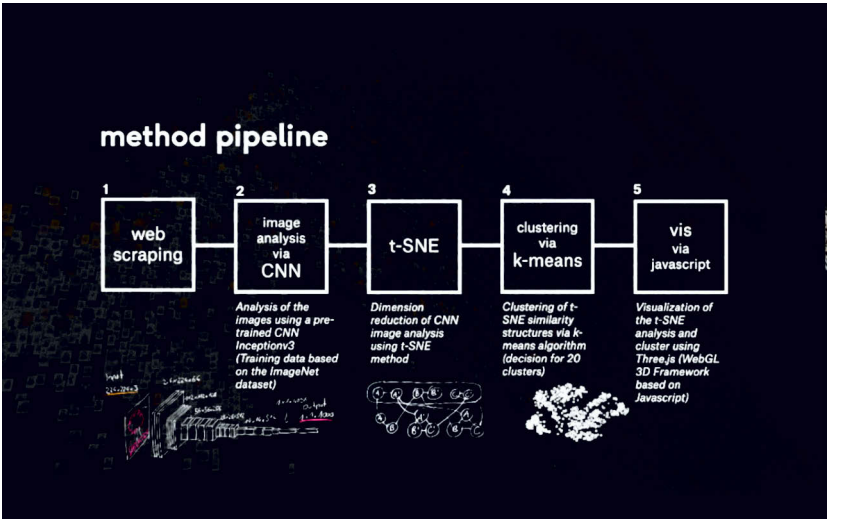
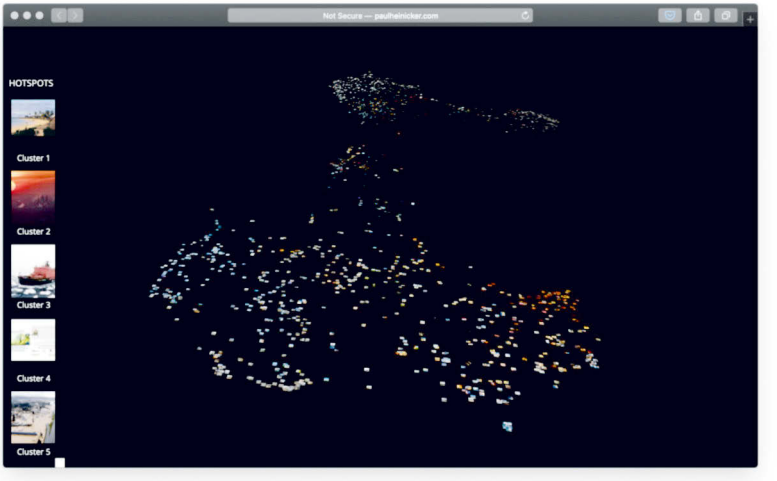


Fig. 5: Screenshot of the interface built on the basis of PixPlot by Yale Digital Humanities Lab. Screenshot, 2020.[18]



18 See Yale DH Lab, 'PixPlot', accessed 19 July, 2021, https://dhlab.yale.edu/projects/pixplot/.

Key motifs and image types

We were able to identify a number of dominant and/or salient motifs and image genres in the scanned images. These motifs were expressed either as technically generated image clusters or as qualitative image groups although differences in the various locales became obvious, too. The key motifs are as follows:

Text-image documents, including
    Text-only documents
    Diagrams
    Covers of books and brochures,
    PowerPoint slides with text,
    Posters with slogans about climate protection
Maps
Infographics and charts
Occasional cartoons or cartoon-style graphics
People, including
    Conferences, group meetings (most of which could be linked to a specific political context)
    Scenes of protest and demonstrations
Depictions of the earth as a globe
Contrasting depictions/diptychs: green landscape versus parched ground
Forest fires
Ice landscape, e.g., polar sea with ice floes
Floods/flooding
Drought/dry soil
CO2 emissions, refineries
Polar bears
Forest/trees/plants
Agriculture/farming/cultivation

Our content review of the image types allowed us to determine the central image types and/or genres within the Google search results. For the most part, three types appeared, sorted by average weighting:

- 1. Photographs, divisible into images depicting (a) the environment, nature, landscape, (b) people (conferences, groups, politics versus protest/demonstrations), (c) the earth, and (d) the polar bear
- 2. Text-image documents, including text-only documents (documents, diagrams, covers of books and brochures, slides with text, activist posters with slogans, maps, cartoons etc.)

- 3. Highly artificial photomontages (representations of contrast of planet earth and landscapes).

Intercultural framing

The t-SNE algorithm serves as a method with which to frame images. The automated image analysis and clustering process of the above-mentioned key motifs, search images and image types helped us to derive statements about visual frames. First is the *frame of vulnerability*, which includes the consequences and impacts of climate change as a problem definition and is the largest image group in all country queries. It includes photographs of icy landscapes (polar sea/glacier/ice floes), drought, flood, heat/forest fire and the polar bear cluster.

Equally dominant and prevalent alongside the frame of vulnerability is the *frame of cause representation*. This was demonstrated in all image queries involving the subject of CO2 emissions (refineries/chimneys with smoke). Photographs of people are also concisely represented; and they can be assigned to the visual *frame of actors*. They symbolise the frame that emphasises approaches to solutions: on the one hand, they tend to be photographs of politicians and international climate conferences, and on the other, they are photographs showing scenes of protest and demonstrations. One particularly interesting result of the Google images was the photographs we were able to assign to the *frame of moral evaluation*. Most significant here were the images relating to the *frame of global anxieties* as embodied in artificially assembled photographs representing the Earth as either the iconic Blue Marble or as a globe or as a burning planet.

Our comparison of the image clusters or groups (via the screenshot table, see Fig. 7) made it possible to determine country-specific distributions as a tendency. The following are our key findings in this realm for the year 2020:

- The t-SNE visualisations of the Google images from Bangladesh and Kenya show the largest number of images featuring people. In particular, they contain scenes of politics, negotiations, conferences and/or meetings. Based on random sampling, the motifs tended to be locally assignable, such as one large set of images representing Prime Minister Sheikh Hasina Wajed of Bangladesh. Surprisingly, very-few-to-no-images featuring politicians could be found in the t-SNE visualisations of the American, Australian and German search queries.
- The Kenyan search query contained the largest proportion of depictions of people engaged in cultivation, followed by Bangladesh and the United Arab Emirates. In an international comparison, Kenyan Google images have the largest proportion of subjects relating to agriculture; and in comparison to the other search queries, this image theme is only marginally represented.

- Also surprising is the marginality of protest images. Images of predominantly young climate protection activists stand for the global use of this image type. Various photos are used 'globally', that is, the same motif appears in different countries. For example, a photo of protesting students and climate activists in New York's Times Square can be found in Google images from Bangladesh and Australia.
- One cluster of polar bears and one of ice landscapes appear throughout every search group. Only in the Kenyan search query is such a cluster missing. Surprisingly, on the other hand, the Kenyan t-SNE visualisation did not show any image cluster relating to the theme of drought.
- With regard to the vulnerability frame, we made an interesting finding about forest images. Internationally, the forest as a motif is predominantly shown to be burning, most clearly in the Google images from the USA, Australia, the United Arab Emirates and Brazil. Only the German Google images showed a group of images that included a technically generated hotspot image depicting the forest as a dried up patch. It would appear that the forest fire in the form of dramatic and high-contrast photographs (black-orange) is superior as a motif of vulnerability to that of drought.
- In general, the theme of heat dominates. The burning globe becomes a denotation of global warming in contrast to the 'Blue Marble'. The t-SNE visualisations, especially those from the USA, Australia and the United Arab Emirates followed by Brazil show a very high proportion of this image motif. In the t-SNE for Kenya and Bangladesh, this image group could not be detected at all.
- A country-specific statement is perhaps identifiable in the group with motifs of flooding. The largest image cluster was located in the search query of Bangladesh, followed by those of the USA and Australia. In particular, photographs featuring people standing in flood waters are found here. Flood scenes from a bird's-eye view, on the other hand, exhibit a global character. For example, photographs showing flooding in Texas after Hurricane Harvey in 2017 are found in Google images from Bangladesh.

Method reflection

The research theme of networked climate images is already embedded in quantitative rules of logic, for example digital image files published and distributed via hypertext transfer protocols on the internet. For this reason, our interests here are more concerned with the degree to which we can have qualitative approaches in digital infrastructures. We want to know how quantitative and qualitative approaches are entangled and how they are differentiated.

Just as our qualitative research is based on established conceptual frameworks, our quantitative analysis is similarly dependent on externally prepared data sources, programming interfaces (APIs) and pre-written algorithms. This is due to the high degree of technical complexity involved in algorithms for image and context analysis and data aggregation on the internet. It means that in most cases it is possible to work solely within a constructed technical framework. This framework, however, is often characterised by services provided by technology and software companies the 'modes of operation' of whose algorithms are difficult to understand.

The *PageRank* algorithm, for example, which orders the search results on Google Images, is responsible for the type and diversity of climate images that we analyse. Around 200 factors influence this algorithm, some of which are known but the most relevant of them are kept well hidden.[19] This is why such a technical framework is often referred to as a 'black box'. Still, it is possible to take a glimpse into the generative system and thus into the construction of the data by using open-source solutions and self-programmed algorithms and the exercise should be part of the analysis.

The qualitative data of our research is based on an analysis of climate images by humanistic means of interpretation. However, the distinction between qualitative and quantitative data suggests a categorically different treatment. Our research experience instead shows that both sides need to be approached with particular scepticism. This is because both share the belief of being artificially constructed and convey—in addition to their values—underlying assumptions and prerequisites. Nevertheless, it is important to note that qualitative data is closer to an anthropocentric understanding of insights whereas quantitative data follows a more structural logic.

In the course of our mixed-methods research, it became clear that it was almost impossible to handle exclusively qualitative data. Instead, qualitative and quantitative approaches interfere with one another in manifold ways. A clear separation between purely qualitative and purely quantitative data could not be maintained. A strict split can be obtained solely on a heuristic level, but it simply doesn't hold true in practice.

Computer-generated image analysis represents the result of various qualitative decision-making processes:

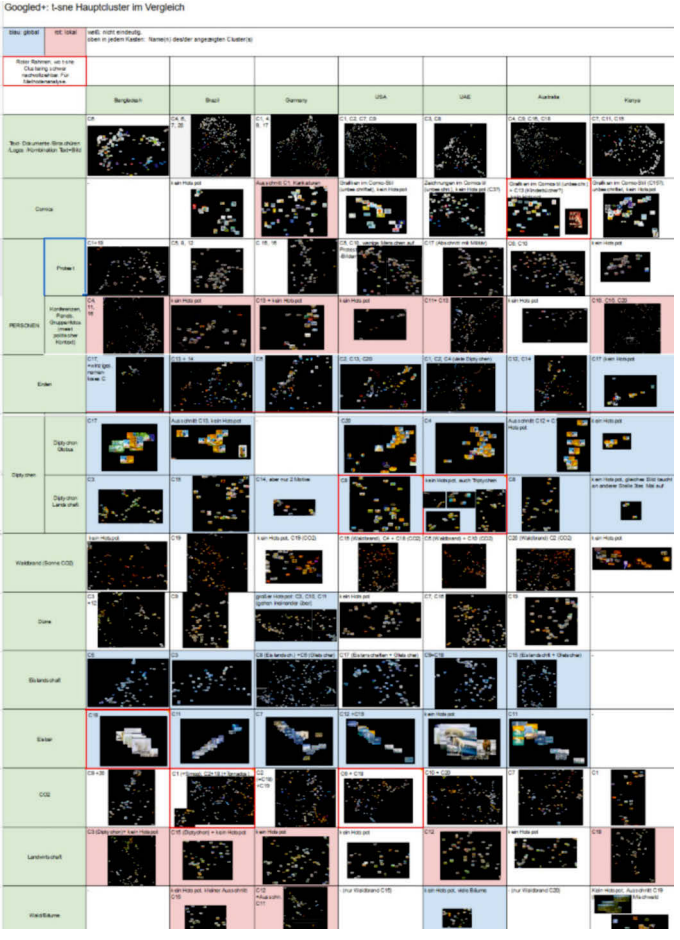
1. The selection of cultural regions to research climate images was undertaken via qualitatively predefined indices.
2. The Google Images corpus itself was based on predefined and partly hidden functionalities of the PageRank algorithm.

---

19 See Danny Sullivan, "Dear Bing, We Have 10,000 Ranking Signals to Your 1,000. Love, Google", in: *Search Engine Land*, accessed July 19, 2021, https://searchengineland.com/bing-10000-ranking-signals-google-55473.

3. The extraction of images was based on a pre-existing framework and network technologies making use of the open-source software Tor to conceal the user's location.
4. Machine learning on inception (pre-trained similarity)
5. We clustered the calculated results that were visualised in a web interface so that groupings became visible to the human eye and could be qualitatively read and used to derive the key motifs, for example in our screenshot table (Fig. 7).

Fig. 7: Screenshot table with image groups of the t-SNE visualisations.



Collaborating in an interdisciplinary team

In the process of our interdisciplinary cooperation, we became increasingly aware of the extent to which humanistic methods of art history and image analysis were themselves quantitative in parts. This is the case as they are focused on image comparison. Conversely, the quantitative method was also subjected to qualitative decision-making processes at various points in the analysis. This occurred in the conception of the algorithm, that is, in the conscious isolation of the image phenomenon from a complex context (experimental arrangement), and in the breaking down of visual phenomena as objects of investigation into formalisable elements (divide and conquer). Even in the adaptation of the data basis and/or the individual functions by, for example, changing the threshold values, the programmer made purely subjective distinctions in consultation with the qualitative question. And, last but not least, the evaluation of the generated results and the subsequent modification of the program in the sense of 'translation' of the image-study analyses also constitute a qualitative-subjective act. By working together, we were forced to reflect intensively on the inadequacy of our respective discipline-oriented views as producing insights that were either solely qualitative or solely quantitative. The group spent a lot of time working to achieve interdisciplinary understanding, which meant we had to explain and make transparent our procedure much more than in homogenous research contexts.

In addition to the process of actual image analysis, we extended our research to include various formats of cross-disciplinary exchange—theoretical as well as practical—by means of experimental formats, such as hackathons and workshops. We also promoted and opened up our research project to a wider audience and tried out various formats in addition to the ordinary research process, similar to labs. For example, we developed open-source tools and held workshops at digital methods institutions.[20]

What is typical for interdisciplinary teams was true for our team too. The members of the group did not work at the same institutes and thus needed to arrange meetings in locations across the campus or, as in our case, across three institutions.

The status of machine learning

Machine learning as a contemporary trend of statistics-led computation was of specific interest to us. However, we were more interested in the actual mathematical models than in the notion of learning, which is predominantly used in superficial

20 See DMI, Climate image spaces and the new climate movements: A Digital Methods Summer school DMI19 project report.

discourses about machine learning that are less interested in its functioning and more in exploiting certain narratives.[21] The use of machine learning calls for a particular case. In our study, we were interested in the automation of a certain idea of image similarity and how it compared to the human understanding of similarity. This resonated with our research scope of testing and reflecting upon a variety of algorithmic approaches. In order to work, a machine-learning algorithm needs to be 'trained' or calibrated preferably with a very large dataset. The 16,000 images we extracted from the Google Images search are too few to be usable for a neural network. For resource reasons, we therefore decided to use a pre-trained neural network following Google's Inception v3 architecture based on the image dataset from ImageNet with over 14 million images. This was a momentous decision considering that ImageNet incorporated taxonomies and categorisations that were not specified for the case of finding similarities in climate images.[22]

We found that machine learning was a challenging but productive method to study large image datasets. In particular, the combination of machine learning and data visualisation offers an alternative to graph-based layout methods. This combination also provides an alternative to purely keyword-based image research where clusters are created by language alone. Instead of looking at a single image artefact in detail or creating a distant view to read patterns of meta-data by using machine learning, we faced a level abstraction through automation. The resulting data dimensions were far beyond the image surfaces at stake in classical image studies and, therefore, a different form of addressing these image structures was needed. Another problem emerged: the choice of parameters determined the distribution of clusters and thus the content reading of the data visualisations. Ultimately, this approach was made productive for qualitative image analysis, as it was based on human interpretation.

Human in the loop—the need for human interaction

When we look at visualisations of machine-learning results, such as the t-SNE visualisations in our international comparison of thousands of climate images, the following question arises: How are these images grouped?

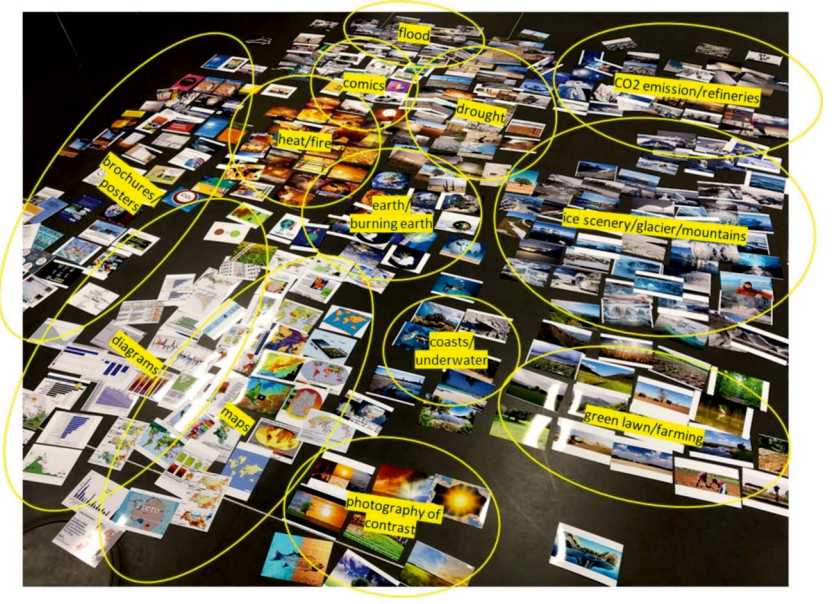
---

21 See Matteo Pasquinelli for similar thoughts. How a Machine Learns and Fails: A Grammar of Error for Artificial Intelligence, 2019, online: https://kim.hfg-karlsruhe.de/how-a-machine-learns-and-fails.

22 In ImageNet, certain types of images are hardly categorized, which is also relevant for our research. The underrepresentation of art is considered in this article: Francis Hunger, "Why so many windows?"– Wie die Bilddatensammlung ImageNet die automatisierte Bilderkennung historischer Bilder beeinflusst, 2021. Zenodo. DOI: https://doi.org/10.5281/zenodo.4742621.

There is no automated method for the recognition of the clusters. We had no other criterion to determine the similarity of the images to one another—other than what we interpreted visually. In decision-making situations, such as boundary-finding and manual clustering, what came into play were our eyes (trained to see climate images), our theoretical knowledge and our ability to interpret. However, the human gaze does not completely reject data visualisations. Surprisingly, human annotations and algorithmic clusters seem to converge to a certain degree in the more discrete groupings (Fig. 8).

Fig. 8: Manual image sorting of the printed out Google search corpus (Germany) with annotations. Photo taken by anci, 2020.



This is significant because the process of creating data visualisations from the source images does not extract semantic information from the dataset. We assume that human annotations also take into account the prominent features of the images, such as shape, colours, contrast and composition when people evaluate the visualisations generated by t-SNE; and they do not merely consider the symbolic meaning as it is not available to the algorithm.

In this respect, the term 'clustering' can prove to be somewhat confusing in methodological approaches. Indeed, it was observed that, in technical terms, the

determination of a t-SNE cluster did not necessarily result from a great number of images, but from a high similarity ratio—most noticeably with duplicate images.[23]

Thus, it can be stated on a technical level that the higher the similarity factor of the images, the more likely it is to be a technically generated image cluster. On the level of human perception, we can say that the higher the frequency of motifs with a common similarity criterion, the more likely it is to be a 'group' of images. (The qualitative determination of it results, in particular, only from a size comparison of all images from all country queries.) The human gaze is thus required for the determination of dominant image groups, as it provides a correction to the algorithmic image sorting. For this reason, we distinguished in our study between the term 'cluster' as a technically generated set of images and the qualitative term 'group' to emphasise image frequency.

The status of data visualisation

In our research, visualisation as a concept and practice is not bound to the idea of depicting the results of the research but rather as an essential element of almost every aspect of our mixed-methods approach. We conceptualised the method of visualisation as a crucial interface for communicating between quantitative and qualitative approaches, that is, between different ideas of how to structure information. On another note, visualisations were used to provide digital data, which would otherwise be incomprehensible to humans, with a subjective form to make them accessible for a qualitative analysis. In our research as an image-based, mixed-methods project in particular, it gave us scope for examining the status of imagery not only as the object of our research but also as a research modus. One interesting gain in knowledge resulting from the data-based image analysis can be found in the ontological image viewing. Indeed, from the point of view of image theory, an interesting translation of image understanding became clear in the mixed use of methods: the digitality of the images enabled constant and iterative changes between them as 1) a phenomenon, as 2) a dataset, as 3) a visualised table and, finally, as 4) a statistical graph.

23 Adrian MacKenzie and Anna Munster clustered images of computer vision research using t-sne in their AI-research to critically reflect on statistical computer vision. See: "Oscilloscopes, slide-rules, and nematodes. Toward heterogenetic perception in/of AI", in: Natasha Lushetich, Iain Campbell (eds.): *Distributed Perception. Resonances and Axiologies* (London: Routledge, 2021), DOI: 10.4324/9781003157021-6.

Discussion

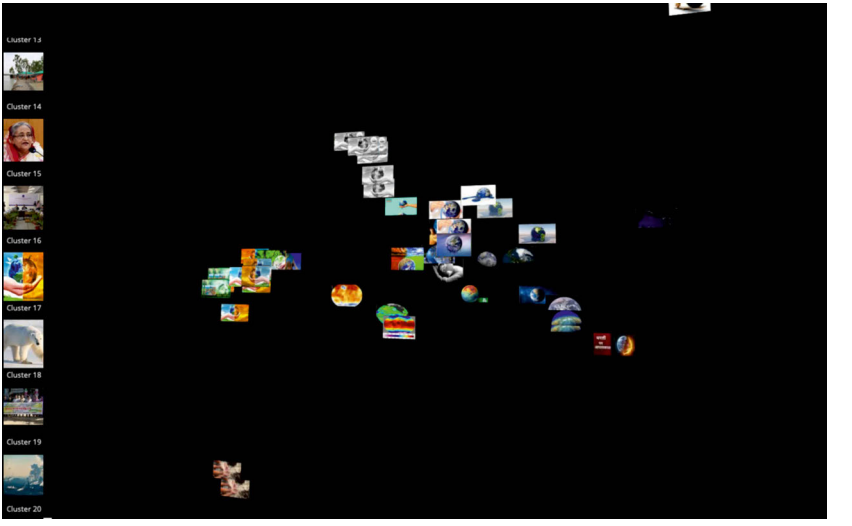
The very idea of the digital as something new and fundamentally transformative—as in the myth of the digital—runs the risk of missing the particularities of what it actually means to work with digital structures.[24] Our focus on digital imagery sharpened our understanding of the algorithmic view. From a qualitative perspective, we increasingly repositioned our research questions and opened them up time and again in the process. The focus on technical methods shifted our emphasis from an interpretative view to a 'structural view', which resulted from algorithmic image recognition. Increased attention to algorithmically detected structures within the image areas and to machine-generated structures of image sorting and clustering became apparent. The analysis of algorithmic cluster structures as the technical 'preliminary work' of image sorting in accordance with the key similarity criteria was also an important component. For instance, the perception of colour contrasts and the dense and heterogeneous distribution of large numbers of images again provided us with ideas and inspiration for new content-related efforts.

One particular conflict in our interdisciplinary cooperation came in the form of a shift of significance regarding the criterion of similarity (tertium comparationis) and the associated research questions. In cooperation with quantitative methods, the qualitatively determined similarity criterion shifted from its content on the phenomenological image level to the level of formalisable data and metadata, the analysis of which is dominated by principles of mathematical structural logic.

This resulted in a change between the image as a phenomenon and the image as a technical dataset. By translating a qualitative question into a regulatory system, certain optimisation problems arose. The focus of our problem resolution was thus entirely on the adaptation of the inner formal logic of the images and thus obscured the view of content-related questions. In other words, causal questions often became questions of quantity and distribution. Differences also emerged in the use of vocabulary regarding the meaning of the term 'cluster' in our intercultural t-SNE-visualisations. The German translations of the word 'cluster' as 'Anhäufung' (accumulation) or 'Gruppe' (group) simply do not fit the technical meaning. With regard to machine-learning approaches, we recommend the thorough use of language so as to provide clarity and avoid confusion; however, we do not recommend referring to the visual clusters in t-SNE as anything other than technically and conceptually feasible (Fig. 9).

24 See Jan Distelmeyer, *Das flexible Kino. Ästhetik und Dispositiv der DVD & Blu-ray / The Flexible Cinema. Aesthetics and Dispositif of the DVD & Blu-ray* (translated using DeepL) (Berlin: Bertz & Fischer, 2012).

Fig. 9: Close-up screenshot from a 'globe' cluster in the t-SNE visualisation.



We need abstraction to grasp details, but details also maintain the models for abstraction. Qualitative methods deal with conceptual abstractions and derive their insights from interpretations of particular structural details as framed by the notion of 'close reading'. Similarly, quantitative approaches look on empirical models and ideas which are then realised as mathematical rule-sets and conventions. In computer science, the paradigm of 'divide and conquer', for example, describes a systemic idea of categorising problems into partial elements.[25] In terms of details, quantitative approaches, especially computer vision and machine-learning algorithms, are good and, if trained properly, prove better than humans at detecting patterns in (media) environments. Therefore, it did not seem like a conflict but rather like an undeniable and more complex relation within the methods themselves. Details can lead to oversimplified assumptions and abstractions can also embrace increased complexities. It is human cognition which always implies reduction since we rely on certain models and world views to adapt to changing realities. From here we can move to a reading of reductionism as a scientific search for causal-deterministic relationships. Our experience gained from a mixed-methods design showed us that a holistic perspective was more appropriate when trans-disciplinary problems were encountered, such as climate-change communication. Overall, the discussion around the idea of reduction was very important to our research, but it could not be properly framed with the two notions of detail and abstraction.

25 See https://en.wikipedia.org/wiki/Divide-and-conquer_algorithm.

Machine processes of image recognition enable new approaches to sorting images according to their formal-stylistic content and structures. The media- and data-critical analysis of the image recognition process of algorithms is decisive in this process. Art-historical classifications in accordance with the style of the images based on the connoisseur's gaze are supplemented by algorithmic glances and technical sorting logics. Both a (phenomenal) image analysis, such as Erwin Panofsky's iconographic-iconological three-phase model of visual levels of meaning, and an algorithmic analysis of the various dataset formats lead to the notion that pictoriality must be understood as the interplay of various image levels.[26] Above all, algorithmic image-sorting helps to manage the sheer mass of images and to find frames. The potential of mixed-methods image analysis can thus be found in its ability to open up a new possibility of analysing images. A more in-depth examination of technical and technological developments, such as machine learning, will lead to a more productive sharpening of the 'statistical gaze'. It does not mean that these technologies of algorithmic image sorting and analysis effectively extend or improve traditional image viewing; they do, however, enable other views that extend an interpretative perception based on conventions with analytical views that result from the media-critical analysis of algorithmic codes.

Uncertainty is a constitutive element of our quantitative and qualitative methods. In contrast to computational models, human models can handle the ambiguous and the undefined. Humanities scholars are aware of the danger, for example, of 'linguistic heteronomy' ('sprachliche Fremdbestimmung')[27] and of individual over-interpretation. The full impact of qualitative considerations reveals itself in the blurry realm of uncertainty. Quantitative approaches also require uncertainty as an essential condition. It tests their limits, thereby making an essential contribution to their productive re-iteration. Dealing with uncertainty is, therefore, indicative of essential mixed-methods ability.

Speaking from a general perspective, the discussion surrounding the success and failure of data-driven research will intensify in such a way that a still higher level of data positivism and "overestimation of statistical significance"[28] will find them-

---

26 See Harald Klinke, "Bildwissenschaft ohne Bildbegriff" / "Image science without image term" (translated using DeepL), in: Harald Klinke, Lars Stamm (eds.): *Bilder der Gegenwart: Aspekte und Perspektiven des digitalen Wandels* (Göttingen: Graphentis, 2013), 11–33.

27 Gottfried Boehm, "Ein Briefwechsel" / "An Exchange of Letters" (translated using DeepL), in: Marius Rimmele et al. (eds.): Bildwissenschaft und Visual Culture (München: Wilhelm Fink, 2014), 26.

28 Hanna Brinkmann; Laura Commare, "Why "Anything goes" der Goldstandard sein sollte – Überlegungen zu Methodentradition und empirischen Forschungsansätzen in den Kunstwissenschaften" / "Why "Anything Goes" Should Be the Gold Standard – Reflections on Methodological Tradition and Empirical Research Approaches in Art Science"(translated using DeepL), in: Verband österreichischer Kunsthistorikerinnen und Kunsthistoriker (VöKK) (ed.): Newest

selves reflected in an increasingly critical manner. With regard to the structure of interdisciplinary DH projects, we assume and hope that there will be an increasing balance of esteem between the humanities and computer-aided disciplines. Compared to the traditional humanities, there is still a focus of interest on computer science as the sole source of technical products and tools.

Nonetheless, the development of qualitative research questions and content is a crucial basis for the design of digital methods and tools. For precisely this reason, a balance should be struck if both approaches are to be considered on an equal footing. Furthermore, future DH endeavours should not be seen as completed, closed-circuit projects that run the risk of ending up in 'DH graveyards'. The goal must be to continuously integrate both into general research and university teaching—the findings of digital imaging and the method-critical approaches from the interdisciplinary interlocking.

Bibliography

Paul Heinicker, Janna Kienbaum, Birgit Schneider, Thomas Nocke: anci website: https://anci.fh-potsdam.de/, 2021.

Boehm, Gottfried, and W.J.T. Mitchell. "Ein Briefwechsel". In *Bildwissenschaft und Visual Culture*, edited by Marius Rimmele et al, 19–40. München: Wilhelm Fink, 2014.

Brinkmann, Hanna, and Laura Commare: "Why 'Anything goes' der Goldstandard sein sollte – Überlegungen zu Methodentradition und empirischen Forschungsansätzen in den Kunstwissenschaften". In *Newest Art History – Wohin geht die jüngste Kunstgeschichte? Tagungsband zur 18. Tagung des Verbandes österreichischer Kunsthistorikerinnen und Kunsthistoriker*, edited by Verband österreichischer Kunsthistorikerinnen und Kunsthistoriker (VöKK), 169–172, 2017. https://voekk.at/sites/default/files/downloads/tagungsbaende/Newest%20Art%20History_VoeKK-Tagungsband.pdf.

Diers, Michael. *Schlagbilder: Zur politischen Ikonographie der Gegenwart*. Frankfurt a. M.: Fischer, 1997.

Dimitrova, Daniela V., and Lulu Rodriguez, Lulu. "The levels of visual framing". In *Journal of Visual Literacy*, 30:1 (2013), 48–65, DOI: 10.1080/23796529.2011.11674684.

Dilly, Heinrich. "Lichtbildprojektion – Prothese der Kunstbetrachtung". In *Kunstwissenschaftliche Untersuchungen des Ulmer Vereins, Verband für Kunst- und Kulturwissenschaften*, edited by Horst Bredekamp et al., 153–172. Gießen: Anabas, 1975.

Distelmeyer, Jan. *Das flexible Kino. Ästhetik und Dispositiv der DVD & Blu-ray*. Berlin: Bertz & Fischer, 2012.

Art History – Wohin geht die jüngste Kunstgeschichte? Tagungsband zur 18. Tagung des Verbandes österreichischer Kunsthistorikerinnen und Kunsthistoriker (2017), 171.

Christ, Katharina et al: DMI, Climate image spaces and the new climate movements: A Digital Methods Summer school DMI19 project report. https://wiki.digitalmethods.net/Dmi/ClimateImageSpaces, 2019.

Entman, Robert. "Framing: Toward Clarification of a Fractured Paradigm". In *Journal of Communication* 43, 4, 1993. https://is.muni.cz/el/1423/podzim2018/POL256/um/Entman_1993_FramingTowardclarificationOfAFracturedParadigm.pdf

Geise, Stephani, and Katharina Lobinger (eds.). *Visual Framing. Perspektiven und Herausforderungen der Visuellen Kommunikationsforschung*, Köln: Halem, 2013.

Grittman, Elke. "Visual Frames – Framing Visuals. Zum Zusammenhang von Diskurs, Frame und Bild in den Medien am Beispiel des Klimawandeldiskurses". In *Visual Framing: Perspektiven und Herausforderungen der visuellen Kommunikationsforschung*, edited by Stephanie Geise and Katharina Lobinger, 95–116. Köln: Halem, 2015.

Hunger, Francis. *"Why so many windows?"– Wie die Bilddatensammlung ImageNet die automatisierte Bilderkennung historischer Bilder beeinflusst,* 2021. Zenodo.

DOI: https://doi.org/10.5281/zenodo.4742621.

Klinke, Harald. "Bildwissenschaft ohne Bildbegriff". In *Bilder der Gegenwart: Aspekte und Perspektiven des digitalen Wandels*, edited by Harald Klinke and Lars Stamm, 11–33. Göttingen: Graphentis, 2013.

Latour, Bruno. *Science in Action. How to Follow Scientists and Engineers Through Society,* Cambridge: Harvard University Press, 1987.

Link, Jürgen. "Literaturanalyse als Interdiskursanalyse: Am Beispiel des Ursprungs literarischer Symbolik in der Kollektivsymbolik". In *Methoden diskursanalytischer Ansätze*, edited by Jürgen Fohrmann and Harro Müller, 284–307. Frankfurt am Main: Suhrkamp, 1988.

MacKenzie, Adrian, and Anna Munster. "Oscilloscopes, slide-rules, and nematodes. Toward heterogenetic perception in/of AI". In *Distributed Perception. Resonances and Axiologies*, edited by Natasha Lushetich and Iain Campbell, 64–81. London: Routledge, 2021, DOI: 10.4324/9781003157021-6.

MacQueen, James B. "Some methods for classification and analysis of multivariate observations". In *Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability*, 1, 1967, 281–297.

Marr, David. *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*, Cambridge, Mass.: MIT Press, 2010 [1982].

Matthes, Jörg. *Framing*, Baden-Baden: Nomos, 2014.

Panofsky, Erwin. *Sinn und Deutung in der bildenden Kunst* [1955]. Köln: DuMont, 1978.

Pasquinelli, Matteo: How a Machine Learns and Fails: A Grammar of Error for Artificial Intelligence, 2019, online: https://kim.hfg-karlsruhe.de/how-a-machine-learns-and-fails.

Schneider, Birgit. *Klimabilder: Eine Genealogie globaler Bildpolitiken von Klima und Klimawandel*. Berlin: Matthes & Seitz, 2018.

Statista. "Number of explicit core search queries powered by search engines in the United States as of April 2021". Accessed July 19, 2021. https://www.statista.com/statistics/265796/us-search-engines-ranked-by-number-of-core-searches/.

Sullivan, Danny. "Dear Bing, We Have 10,000 Ranking Signals to Your 1,000. Love, Google". In *Search Engine Land* Accessed July 19, 2021. https://searchengineland.com/bing-10000-ranking-signals-google-55473.

Thürlemann, Felix. "Bild gegen Bild: Für eine Theorie des Vergleichenden Sehens". In *Zwischen Literatur und Anthropologie – Diskurse, Medien, Performanzen*, edited by Aleida Assmann, Ulrich Gaier and Gisela Trommsdorff, 163– 174. Tübingen: Narr, 2005.

Hinton, Geoffrey and Laurens Van der Maaten. *Visualizing Data using t-SNE*, 2008, accessible at https://jmlr.org/papers/volume9/vandermaaten08a/vandermaaten08a.pdf

Walsh, Lynda. "'Tricks', Hockey Sticks, and the Myth of Natural Inscription: How the Visual Rhetoric of Climategate Conflated Climate with Character". In *Image Politics of Climate Change Visualizations, Imaginations, Documentations*, edited by Thomas Nocke and Birgit Schneider, 81–105. Bielefeld: Transcript, 2014.

Walsh, Lynda. "TheVisual Rhetoric of Climate Change". In: *Wires Climate Change*, Volume 6, 4, (2015), 361–368.

Warburg, Aby. *Der Bilderatlas Mnemosyne*. Berlin: Akademie Verlag, 2008.

Warburg, Aby. "Einleitung zum Mnemosyne-Bildatlas [1929]". In: *Kulturwissenschaft. Eine Auswahl grundlegender Texte*, edited by Uwe Wirth, 137–145. Frankfurt am Main: Suhrkamp, 2008.

Wikipedia: Divide-and-conquer-algorithm. Accessed December 2022 at https://en.wikipedia.org/wiki/Divide-and-conquer_algorithm.

Yale DH Lab. "PixPlot". Accessed July 19, 2021.https://dhlab.yale.edu/projects/pixplot/.

#QUANTIFICATION

The Oxford English Dictionary understands #Quantification as a terminus in the realm of logic: "the ascription of universal or particular quantity to a proposition, or, more usually, a term. In modern Logic, Mathematics, and Linguistics: the use of a quantifier to indicate the scope of a variable, etc." as well as the "action of quantifying something; an instance of this." *

The mixed methods projects underline #Quantification as a crucial process in digital humanities, linking the spheres of humanities into the structures of computer science by "transforming real world observations and phenomena into a numerical structure. It is both an abstraction in the sense of reducing a particular quality for a specific operational purpose and a projection in terms of re-framing something in a new context" (ANCI). The foci in detail differ and sketch a broader meaning of #Quantification.

One looks at the specifics of the actual 'translation' of qualitative parameters embodied in a #Corpus by means of annotation "as a way of making vaguely defined theories and hypotheses concrete" (BachBeatles, see #Human-In-the-loop) Another focuses at the ensuing findings and their qualifying characteristics (Handwriting). Yet other projects underline the values of a parallel analysis of quantitative corpora for a deeper understanding (Rhythmicalizer), and come full circle to see quantification "as a complement to the qualitative part, which may tackle questions not addressed (and not addressable) in the statistical approach" (BachBeatles, ArchiMediaL).

In result, #Quantification emerges not only as a terminus but also as one of the key concepts of digital humanities and probably even more so for projects mixing methods (ArchiMediaL).

*"Quantification, n.". in: *Oxford English Dictionary (OED)*, Third Edition, December 2007; most recently modified version published online March 2022, https://www.oed.com/ [accessed: 25.10.2022].

Title: Tracing patterns of contact and change: Philological vs. computational approaches to the handwritings of a 18th century migrant community in Berlin
Team: Humboldt-Universität zu Berlin: Prof. Dr. Roland Meyer (Primary investigator); Aleksej Tikhonov, M.A.; Ewa Kolbik, B.A.; Dr. Robert Hammel (Consultant) – Fraunhofer IPK Berlin: Dr.-Ing. Bertram Nickolay (Primary investigator); Dr. Jan Schneider (Consultant) – MusterFabrik Berlin: Klaus Müller, M.Sc.; Dipl.-Ing. Maxim Schaubert; Luisa Esguerra, M.Sc.; Dr. Marc von der Linden (Consultant) – Archive in the "Bohemian village", Berlin: Stefan Butt (Consultant)
Corpus: Digitized handwritten personal vitae and sermon manuscripts, 18th-early 19th c., from the Archive in the "Bohemian village", Berlin – ca. 5000 pages
Field of Study: (Historical) philology, palaeography, Czech linguistics, language contact; digital restoration of documents; image recognition, authorship attribution, handwriting recognition
Institution: Humboldt-Universität zu Berlin, Fraunhofer IPK Berlin, MusterFabrik Berlin
Methods: High-quality document scanning (own developments of Fraunhofer IPK and MusterFabrik); Image recognition and analysis, Optical character recognition, Machine-learning, Neural networks; Linguistic analysis, historical morphology, Annotation of document images
Tools: Assistance system LiViTo (own development)
Technology: Supervised machine-learning, neural networks, Python programming, tagging, corpus linguistics

Detecting Authorship, Hands, and Corrections in Historical Manuscripts. A Mixed-methods Approach towards the Unpublished Writings of an 18th Century Czech Emigré Community in Berlin (Handwriting)

Roland Meyer, Aleksej Tikhonov, Robert Hammel

Abstract *When one starts working philologically with historical manuscripts, one faces important first questions involving authorship, writers' hands and the history of document transmission. These issues are especially thorny with documents remaining outside the established canon, such as private manuscripts, about which we have very restricted text-external information. In this area – so we argue – it is especially fruitful to employ a mixed-methods approach, combining tailored automatic methods from image recognition/analysis with philological and linguistic knowledge. While image analysis captures writers' hands, linguistic/philological research mainly addresses textual authorship; the two cross-fertilize and obtain a coherent interpretation which may then be evaluated against the available text-external historical evidence. Departing from our 'lab case', which is a corpus of unedited Czech manuscripts from the archive of a small 18th century migrant community, the Herrnhuter Brüdergemeine (Brethren parish) in Berlin-Neukölln, our project has developed an assistance system which aids philologists in working with digitized (scanned) hand-written historical sources. We present its application and discuss its general potential and methodological implications.*

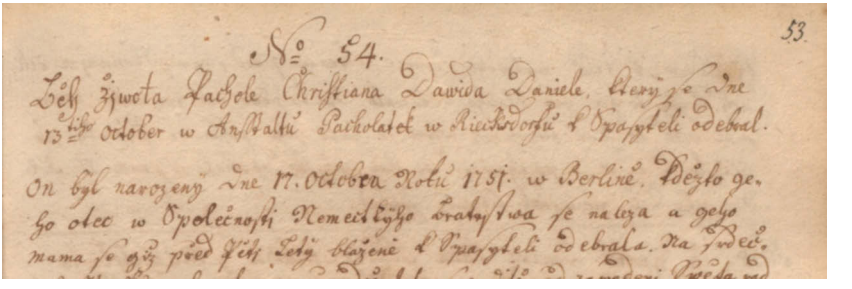
Project description

For all humanities, historical manuscripts are an essential source of knowledge. Interest in textual history has gone so far as to make philologists fear that textual content may become backgrounded in comparison to issues of textual genesis, versions and manuscripts. This fear is however unfounded, as long as researchers are aware of versions of texts and do not rely only on apparent originals or *urtexts*.[1]

1 Livia Kleinwächter, "The Literary Manuscript: A Challenge for Philological Knowledge Production", in: Pál Kelemen and Nicolas Pethes (Ed.): *Philology in the Making Vol. 1* (Bielefeld: transcript Verlag, 2019), 109–128.

Traditional philology and other text-centred humanities have developed a received methodology of accessing old manuscripts which involves research on the text-external context, close reading, transcription, critical edition and time-consuming textological 'detective work'. On the other hand, modern image and pattern recognition techniques promise to be able to distinguish personal handwritings and isolate pre-defined graphic templates in them in an automatic (supervised) way. The present project aimed at confronting these two methodologies, reflecting systematically upon the question as to which of the two approaches was more adequate and successful, and combining their benefits. While large collections of handwritten texts produced by different unknown scribes pose a great challenge, they are also an ideal testing ground for applying new research methods which unify computational and linguistic approaches. It is instrumental here to distinguish between the 'scribe' as the material producer of the manuscript at hand and the 'author' as its intellectual originator. The author either delivered the original version of the text on which the manuscript is based, or (s)he dictated its contents to the scribe. Consequently, optically recognizable handwriting features of the manuscript can be attributed to the scribe whereas linguistic features should rather be ascribed to its author. The manuscripts of 18th c. Czech Protestant immigrants in Prussia are an example of such collections of texts. The manuscripts, which include autobiographies of parishioners of the Moravian Church (*Herrnhuter Brüdergemeine*) as well as a large number of so-called Choir speeches (a type of sermons typical for that Church), are written in Czech using *Kurrent* script then common not only among Germans but also among Czech speakers.

Fig. 1: Excerpt from a biography manuscript from the Rixdorf archive. The headline script differs slightly from the body text, as it probably had a decorative function.



The manuscripts are kept in the archive of the *Herrnhuter Brüdergemeine* in Berlin-Neukölln. This material object of study consists of about 5,000 hand-written pages in (historical) Czech, dating from about 1740 to 1830, from a small community of re-

ligiously persecuted migrants (called *exulants*) to Berlin, the ancestors and founders of the present-day Moravian Church Parish.

The Czech immigrants, originally adherents to the Protestant Unity of Brethren, escaped from Catholic counter-reformation in Bohemia and Moravia and, after a many-year odyssey, were eventually accepted in Prussia by King Frederick William I. In 1737, some of them settled in the small village of Rixdorf, then on the outskirts of Berlin but now part of it. In 1756, most of the immigrants joined Count Nikolaus Ludwig Zinzendorf's Moravian Church.[2] The Czech language was commonly used by these immigrants at least until the early 19th c. before it was gradually replaced by German. Surrounded by a German-speaking environment, Czech speakers in Rixdorf thus formed a particularly interesting language enclave during a period when the language in the Czech mainland was suffering a considerable decline.

Our project focused on the autobiographies of Rixdorf parishioners as these documents were of special historical and linguistic interest. Up to the present day, writing an autobiography is part of the religious duties of every member of the Moravian Church. The autobiography is supposed to cover important stages of the parishioner's life with particular emphasis on spiritual aspects.[3] We initially supposed that processes similar to those described by Mettele also applied to Czech immigrants' autobiographies; parishioners with little practice in using the Czech written language, such as peasants and craftsmen, probably needed the assistance of educated authors who would turn their oral accounts into written texts. The texts were later copied and corrected, the latter being evident especially from numerous subsequent amendments.[4] The text corpus comprises 183 autobiographies covering a period between 1760 and 1819 with a total number of 660 handwritten pages.[5] A selection of these were published in abridged form by E. Štěříková in modern Czech orthography[6] and later also translated into German.[7] Unfortunately, the autobiographies contain no explicit indication of their authors or scribes. For the community, it is important to uncover the content of these texts, the people who were able or authorized to write them, and the history of their transmission. Given the influence of

2 For an account of the history of Czech settlement in Rixdorf see Manfred Motel, *Das böhmische Dorf in Berlin: die Geschichte eines Phänomens* (Berlin: Darge Verlag, 1983).

3 On German brethren autobiographies see Gisela Mettele, *Weltbürgertum oder Gottesreich: die Herrnhuter Brüdergemeine als globale Gemeinschaft 1727–1857* (Göttingen: Vandenhoeck & Ruprecht, 2009).

4 For more information on the manuscripts held at the Rixdorf archive see Aleksej Tikhonov, *Sprachen der Exilgemeinde in Rixdorf (Berlin): Autorenidentifikation und linguistische Merkmale anhand von tschechischen Manuskripten aus dem 18./19. Jahrhundert* (Heidelberg: Winter Verlag, 2022), 83–97.

5 Tikhonov, *Sprachen*, 58.

6 Edita Štěříková, *Běh života českých emigrantů v Berlíně v 18. století* (Praha: Kalich, 1999).

7 Edita Sterik, *Die böhmischen Exulanten in Berlin* (Herrnhut: Herrnhuter Verlag, 2016).

the Herrnhuter Brüdergemeine on the Lutheran Church and that of the *exulant* communities on the Berlin city history, the record of linguistic and cultural adaptation implicit in the texts earns a broader general interest.

Given the various participants involved in the completion of an autobiography, a major goal of the project was to determine the number of different authors and scribes engaged in it, and thus to reconstruct the history of the manuscript. Crucial clues to the reconstruction are provided by linguistic features of the autobiographies, on the one hand, and by visual features of the handwritings on the other. The twofold analysis of both types of features revealed that the 183 autobiographies had been produced by a total of 26 different authors and 12 different scribes. The results of the research project are summarized in Aleksej Tikhonov's PhD thesis[8] and in a number of recently published papers.[9]

Moreover, an open-source software tool called LiViTo[10] was developed to provide an assistance system for the analysis of historical manuscripts. The tool comprises modules for scribe and keyword detection as well as modules for revision detection and linguistic feature analysis.[11] Researchers from both teams – linguists and engineers – jointly developed the tool. It is language-independent and was published as adaptable open-source software in order to make it useable beyond the 'lab case' addressed in the project. An unexpected achievement was the rapid improvement made in the optical character recognition (OCR) of historical individualized handwriting, by using machine-learning techniques with neural networks. Thus, we can now actually search textually in the digitalized document images and identify repeated occurrences of keywords. Finally, based also on neural network technology, LiViTo is able to find various types of corrections and amendments by detecting layers of handwriting on the basis of image processing. This helps linguists to group

8 Aleksej Tikhonov, *Autorenidentifikation und linguistische Merkmale der Rixdorfer Handschriften: Eine Untersuchung anhand von Manuskripten aus dem 18./19. Jahrhundert* (Dissertation) (Berlin: Humboldt-Universität zu Berlin, 2020). Tikhonov, Sprachen (2022).

9 Aleksej Tikhonov and Klaus Müller, „LiViTo: A software tool to assess linguistic and visual features of handwritten texts", in Adrian Paschke, Clemens Neudecker, Georg Rehm, Jamal Al Qundus, Lydia Pintscher (Ed.): *Qurator* – Conference on Digital Curation Technologies *2020* (Berlin: Online-Open-Access-Publication, 2020), https://ceur-ws.org/Vol-2535/paper_8.pdf.

Klaus Müller, Aleksej Tikhonov, Roland Meyer, „LiViTo: Linguistic and Visual Features Tool for Assisted Analysis of Historic Manuscripts", in: *Proceedings of the Twelfth Language Resources and Evaluation Conference* (Marseille: European Language Resources Association, 2020), 885–890.

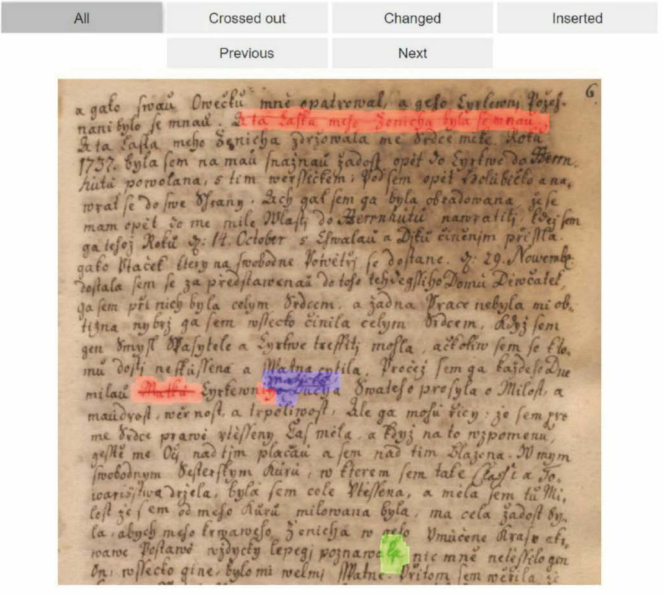
Aleksej Tikhonov and Klaus Müller, „Scribe versus authorship attribution and clustering in historic Czech manuscripts: a case study with visual and linguistic features", in: *Digital Scholarship in the Humanities* 37 (Oxford: Oxford University Press, 2022), 254–263.

10 Tikhonov and Müller, *LiViTo* (2020).

11 Tikhonov and Müller, *LiViTo*, 885–890.

texts by potential later correctors and form hypotheses about their identity; conversely, linguists' classifications provide training data for the refinement of the image processing component.

Fig. 2: An example of LiViTo's revision detection function.



The present paper provides an outline of the methods applied to the analysis of the manuscripts and a discussion of the results.

Both methodologically and content-wise, the project has turned out extremely fruitful. After high-quality digitization of the documents, both teams used their respective methods to add information that could help to identify scribes and/or original authors: annotation of specific linguistic properties for team (i) and graphics-based machine-learning techniques for team (ii). Both approaches were systematically examined during regular weekly common work sessions, and led to a mutual refinement of the methodology (e.g., as to which parts of the script were distinctive) and to a deeper understanding of the respective results. An interesting and unexpected outcome was that only for part of the autobiographies was there strong agreement between linguistics-based and graphics-based classifications. Another set of texts, however, was considered diverse by the linguists, but homogeneous by the image processing group; the obvious explanation was that these texts had been written up by one scribe or copied from the original sources later. Clearly, none of the

two approaches could have achieved this result without the other—both necessarily complement each other in detecting document history. However, it also proved important in the final phase of the project to confront both findings with historical background knowledge from the archives in order to achieve a sound explanation.

State of related research

Since the present project combines various scientific methods and disciplines, current research must be taken into account in at least three:[12] (i) linguistic and visual author and scribe attribution, (ii) stylometric research and (ii) computer aided keyword analysis/search in digital documents. Burrows designed a method of analysing word frequencies to visualize the distance between two or more texts in terms of authorship.[13] Another comparable measurement is Kullback-Leibler divergence (KL divergence). KL divergence is of greater importance because it is not based solely on the relationship between individual word frequencies, but on the stochastic Markov chain and the probability distance.[14] The central role of function words in multivariate analysis is implemented in machine learning approaches in which text categorization is based on neural networks. This method of author and scribe assignment has been widely used in various disciplines since 1993.[15] Hope[16] studied the authorship of Shakespeare's plays, exploring the connections between John Fletcher, Thomas Middleton, and Shakespeare.[17] The R package stylo, developed by Eder, Ry-

12 For a detailed overview: Tikhonov and Müller, *LiViTo* (2020); Müller et al., *LiViTo*, 885–890; Tikhonov and Müller, *Scribe*, 254–263.

13 John F. Burrows, "Word-Patterns and Story-Shapes: The Statistical Analysis of Narrative Style", in: *Literary and Linguistic Computing* 2 (Oxford: Oxford University Press, 1987), 61–70. John F. Burrows, "'An Ocean Where Each Kind…': Statistical Analysis and Some Major Determinants of Literary Style", in: *Computers and the Humanities* 23 (New York/Heidelberg/AA Dordrecht: Springer, 1989), 309–321.

14 Moshe Koppel, Jonathan Schler, Shlomo Argamon, "Computational Methods in Authorship Attribution", in: Steven Sawyer (Ed.): *JASIST* 60 (Hoboken: Wiley-Blackwell, 2009), 9–26.

15 Koppel et al., *Computational*, 11.

16 Jonathan Hope, *The authorship of Shakespeare's plays. A socio-linguistic study* (Cambridge: Cambridge University Press, 1994).

17 For a recent statistical account see also Petr Plecháč, „Relative contributions of Shakespeare and Fletcher in Henry VIII: An analysis based on most frequent words and most frequent rhythmic patterns", in: *Digital Scholarship in the Humanities* 36 (Oxford: Oxford University Press, 2021), 430–438.

bicki & Kestemont[18] is a tool for statistical analysis of the style of one or more texts. In recent years, stylometric techniques in combination with 'stylo' have become popular among scholars in humanities who are concerned with the question of authorship of texts and with language statistics.[19] The use of tools for authorship analysis needs digital input data, but as most historical documents are not digitized and the manual transcription process itself is very time consuming, there has been considerable research on automatic optical character recognition (OCR). One of the first systems capable of transcribing more than just single well separated characters was the omni-font software developed by Kurzweil Computer Products in 1974.[20] A prominent free open-source tool for OCR, which can transcribe various languages and styles is Tesseract.[21] Recent development in machine learning led to first research results on algorithmic handwritten text recognition (HTR), which are on human level accuracy.[22] Inspired by these technological improvements Transkribus, a service platform for computer-aided transcription, was developed in 2017.[23]

Method reflection

Participants of the project, prerequisites

The project was jointly headed by Roland Meyer, chair of West Slavic languages at Humboldt University (HU) in Berlin, and Bertram Nickolay of Fraunhofer Institute

18 Maciej Eder, "Does Size Matter? Authorship Attribution, Small Samples, Big Problem", in: *Digital Scholarship in the Humanities* 30 (Oxford: Oxford University Press, 2010), 167–182. Maciej Eder, Jan Rybicki, Mike Kestemont, „Stylometry with R: a package for computational text analysis", in: *R Journal* 8 (1), (Online-Open-Access-Publication, 2016), 107–121.

19 See the stylometric analysis of direct speech in the television series *The Big Bang Theory*: Maryka van Zyl and Yolande Botha, "Stylometry and Characterisation in The Big Bang Theory", in: *Literator* 37/ 2 (Cape Town: Aosis Publishing, 2016), 1–11.

20 J. Scott Hauger. *Reading Machines for the Blind: A Study of Federally Supported Technology Development and Innovation* (Dissertation) (Blacksburg: Virginia Polytechnic Institute and State University, 1995).

21 Anthony Kay, "Tesseract: An Open-Source Optical Character Recognition Engine", in: *Linux Journal*, (Online-Open-Access-Publication, 2007).

22 Alex Graves, Santiago Fernández, Faustino Gomez, Jürgen Schmidhuber, "Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks", in: *Proceedings of the 23rd International Conference on Machine Learning* (Pittsburgh: Carnegie Mellon University, 2006), 369–376.

23 Philip Kahle, Sebastian Colutto, Günter Hackl, Günter Mühlberger. "Transkribus – A Service Platform for Transcription, Recognition and Retrieval of Historical Documents", in: *14th IAPR International Conference on Document Analysis and Recognition (*ICDAR) (Kyoto: IEEE, 2017), 19–24.

for Production Systems and Design Technology (Fraunhofer IPK) in Berlin. A team under Bertram Nickolay is known to have designed an efficient technology and software assistance system for piecing together torn and shredded paper documents of former East German State Security. Musterfabrik Ltd, a company affiliated with Fraunhofer IPK, continues the digital reconstruction of (two-dimensional) cultural assets, including, for example, the written remains of the recently collapsed Cologne city archive, or the fragmented hand-written notes of G.W. Leibniz.[24] Klaus Müller of Musterfabrik Ltd mainly carried out the research for the part of the present project involving optical pattern recognition. He was accompanied by Maxim Schaubert and head of Musterfabrik Marc von der Linden (consultant). As a prerequisite for the project, all of the Czech manuscripts kept at the archive in Berlin-Neukölln were scanned at Musterfabrik by Luisa Esguerra Rodriguez with an overhead scanner using a resolution of 400dpi and a bit depth of 24-bit colour. The quality of the scans proved sufficient for a computational analysis of handwriting features.

Linguistic research for the project was conducted by the team at HU, which has a strong background in Czech (historical) linguistics and in corpus linguistics. The research was undertaken primarily by Aleksej Tikhonov with the assistance of Ewa Kołbik. Slavic and computational linguists Roland Meyer and Robert Hammel regularly contributed their expertise and acted as supervisors. There was a close exchange both during the preparation and training of models of visual variation, and during statistical analysis across the teams.

An absolutely essential ingredient of the research was cooperation with the archive of the *Herrnhuter Brüdergemeine* in Berlin-Neukölln and with *Archiv im Böhmischen Dorf*, headed by Stefan Butt. Butt generously provided advice and orientation in Brethren traditions; and the *Brüdergemeine* parish kindly made available their manuscripts for digitization, handwriting recognition, and analysis of authorship. The project remunerated them with professional digital preservation of their manuscripts, archival contract research, joint outreach activities and, last but not least, unlocking of the contents of the documents which is very important for the community's historical record.

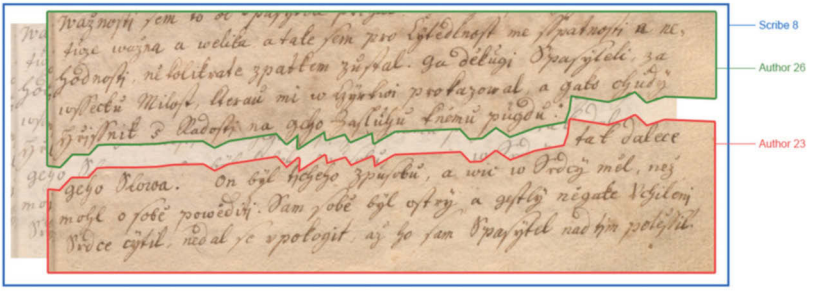
A. Quantification

Our initial goal was to match and align large quantities of linguistic and visual data in order to identify authors and scribes in our corpus of 18th c. Czech Brethren autobiographies. As already mentioned, we departed from the assumption that the production of the autobiographies involved at least two more parts than that of the oral

24 "Analyse der ‚Rixdorfer-Predigten,'" MusterFabrik Berlin, accessed February 2, 2023, https://musterfabrik-berlin.de/landingpage/index.php/rixdorfer-predigten/.

autobiographical account itself, namely, the part of the author as producer of a coherent text and the part of the scribe as producer of the manuscript, the latter above all imposing characteristic orthographic features on the text.

Fig. 3: Example of the distinction between author and scribe.



Mixed methods in the case of the Rixdorf autobiographies thus require a well-defined set of linguistic features believed to sufficiently characterize authors' languages, a set of distinctive orthographic features of scribes,[25] and a set of visual handwriting features which can distinguish scribes.[26] Both approaches involve mathematical methods of determining the similarity between different manuscripts by way of data clustering. Data visualization is used to present the distances between data clusters of similar manuscripts (see G below). In the case of linguistic authorship and scribe identification, linguistic stylometry provides well-established quantitative methods and even appropriate software packages. The present project mostly relied on the software package Stylo.[27] In the case of optical pattern recognition, a computer-aided analysis is the only possible method of coping with a large quantity of data since retrieving similar handwriting features from a large corpus of texts is well beyond the limits of manual analysis.

B. Qualitative data from a linguistic and from a visual perspective

In the present project, the identification of authors and scribes is based on a simultaneous analysis of two qualitatively different data sets, that is linguistic (including orthographic) and visual data. The linguistic features entering the analysis include morphological and syntactic parameters such as particular noun desinences, different infinitive forms, average sentence length, the omission of subject pronouns

25 Tikhonov, *Sprachen*, 137–155.

26 Tikhonov & Müller, *LiViTo* (2020).

27 Eder et al., *Stylometry* (2016).

('pro-drop'), colloquial vs. literary lexical elements etc. Orthographic features taken into consideration comprise the use of particular orthographic systems, different spellings of geographical names and also the presence of various types of revisions such as visibly marked deletions or additions to the manuscript.

The linguistic and orthographic similarity of different manuscripts was calculated based on Euclidean distance between the feature vectors. Both types of qualitative data allow a classification of the manuscripts at hand according to how many authors and scribes were involved in their production.

There is, however, a heuristic difference between the two types of data sets. While the linguistic and orthographic features used in the stylometric analysis of the manuscripts were deliberately chosen by the researcher on the basis of his knowledge of Czech language history, the optical pattern recognition rests on an analysis of no less than 128 different visual handwriting features automatically chosen by the computer program.[28] In comparison, in her handbook of forensic handwriting analysis Seibt[29] discusses only 60 different characteristic features of individual handwriting that should be noticed by examiners when they compare documents. These include, for example, pen pressure, beginning and end strokes, spacing between words etc. The present research on the Rixdorf manuscripts took into account more than twice as many visual features. A final, truly qualitative source of data for the project were historical records about the Brethren in research literature, which allowed Tikhonov (2020) to finally ascribe most of the identified anonymous authors and scribes plausibly to actual historical persons.

C. Uncertainty

Not only do stylometric linguistic feature analysis and optical pattern recognition require different models to interpret the data, but both models have also to be eventually merged in order to develop a unified picture of distance and similarity between the different manuscripts. It turned out that stylometric linguistic analysis and semi-automatic optical pattern recognition of handwriting did not always produce identical results, so researchers had to clarify the fuzziness between the results of both analyses. This was accomplished in the following manner:

In the initial phase of the project both linguists and computer scientists defined their own sets of potentially relevant features. While visual handwriting features were obtained automatically by Musterfabrik Ltd software, characteristic linguistic features, including orthographic ones, were devised by the researchers and subsequently tested on a small sample of texts. Preliminary results of both approaches

28 Müller et al., *LiViTo*, 887.

29 Angelika Seibt, *Unterschriften und Testamente – Praxis der forensischen Schriftuntersuchung* (München: Beck, 2008), 97–142.

were then compared. While the linguistic analysis yielded 12 subclusters of similar manuscripts equalling 12 different potential authors/scribes, optical pattern recognition of handwriting features resulted in only 10 different subclusters (scribes).

A close comparison of both results revealed that 10 out of the 12 'linguistic' subclusters essentially matched the subclusters identified by visual pattern recognition. However, a number of texts, which were assigned to different scribes by the two approaches, were in fact at the statistical boundary between two separate subclusters and thus could belong to either of two scribes. Finally, 9 out of 12 scribes could be plausibly identified with historical persons, whereas three scribes remain either controversial or completely unknown. The corresponding subclusters may be classified as hybrid since they do not allow unequivocal identification of author or scribe.

D. Interpretable models

Computational (machine-learning) methods and linguistic/stylometric methods generally focus on different aspects of our research question: scribe detection based on visual features, on the one hand, and authorship attribution based on features of language and style on the other. However, there is also an overlap, especially when (computational) word or grapheme detection or revision detection assist linguistic analysis.[30] Both visual pattern recognition and linguistic/stylometric analysis start out with sets of features which function as vectors or dimensions along which texts vary. In the case of visual patterns these features are machine-learned, but in the linguistic/stylometric case they are deliberately chosen and annotated.

A dimension reduction technique (T-distributed Stochastic Neighbour Embedding, t-SNE) is applied in order to visualize the clustering of texts in a 3-dimensional space. The clusters are then interpreted as texts belonging to the same scribe; this concludes the modelling of visual patterns.

Linguistic/stylometric modelling also starts by clustering, but then continues by qualitative analysis of many aspects of the manuscripts, from inspection of single features to historical background knowledge about the persons involved. The linguistic characteristics which form the basis of the clustering are often immediately interpretable. For example, certain endings of words point to a colloquial rather than formal register. Certain word orders (e.g., verb-final in embedded clauses) or a relatively low frequency of null subjects and a high amount of third person subject pronouns would point to German influence. In other cases, dimensions of variation 'just work' in distinguishing individual styles, but a comprehensive interpretation is hard to devise; this would apply, for example, to certain spellings of names or to measures such as average sentence length. In any event, the interpretation of the *model* of authorship and document transmission essentially involves sets of triples of text,

30 Müller et al., *LiViTo* (2020).

author and probability of authorship; but in many cases it also involves individual histories of rewriting and copying.

E. The status of machine learning

While the analysis of linguistic and orthographic features is done more or less manually, the optical pattern recognition technique mainly relies on machine-learning algorithms. Machine-based detection of similar visual handwriting features requires preliminary training based on limited samples of at least five pages from two distinct scribes, that is, about 10 pages of handwritten text.

It is not yet clearly understood which handwriting features are selected by the computer program in the course of training for detecting similarities between different handwritings. Machine learning thus effectively replaces a process of forensic handwriting analysis which relies on a smaller set of features, careful attention, and knowledge by experience, but reaches its limits when it comes to large collections of unknown sources. At the same time, it is clear that due to the complexity of text production—potentially involving distinct autobiographic reporters, authors, scribes, later copyists and correctors —, machine learning of handwriting differences can only contribute partially to the actual research issue of document histories. It must be integrated with independent stylometric/linguistic, textological and historical knowledge, calling for a mixed-methods approach.

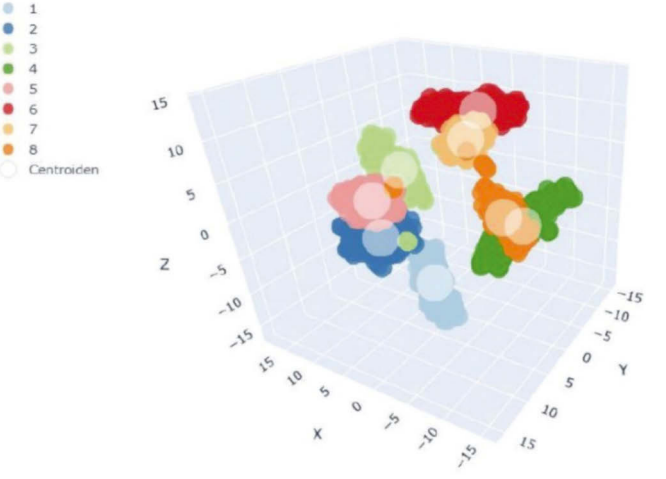
F. The 'human in the loop'

Human intervention is necessary at many points in the workflow: The user initially uploads manually transcribed texts in order to form a ground truth. LiViTo uses transliterations to train itself for the particular case. The user then has to form hypotheses about the number of possible scribes by uploading automatically created line segmentations that probably belong together into the system folders (exact instructions are available on Gitlab).

The statistical methods of stylometry leave many parameters to be determined by the researcher (including, e.g., the choice of distance measure, clustering method, or set of most frequent words); the selection of linguistic features is based on philological wisdom; and historical aspects are investigated by classical rather than digital methods. The classic methods would be for example, research into personalities who were able to write in the community, reconstruction of the history of the handwriting, formulation of possible educational paths in the community. Similarly, the process of machine-based optical pattern recognition involves several steps of manual control during which the recognition process is halted and intermediate results are checked and possibly corrected along the way.

Tikhonov's[31] method exemplifies this procedure: Initially, the user estimates a number of potential writers and manually transcribes a sample of the manuscripts, consisting of at least five pages or 100 lines per potential writer. LiViTo trains a neural network and transcribes further texts from the examined sample. The network architecture for the transcription network is a CNN-LSTM-CTC. Outputs from the convolutional neural network (CNN) are fed into a special form of a recurrent neural network, a long short-term memory (LSTM) network designed to handle temporal data structures. The connectionist temporal classification (CTC) function then interprets the sequence of the LSTM outputs as a probability distribution over all possible transcriptions for a given input sequence and trains the network by maximizing the log probabilities of the correct transcriptions on the training set.[32] The scribe identification network achieved an identification accuracy of 85% on our dataset. To make the results human-readable and interpretable, we took the network's output and embedded the 128 automatically chosen visual features to get a three-dimensional vector.

Fig. 4: Visually based writer clustering in LiViTo.



This vector results from several loops of exchange between the user and the machine. During the development phase, these were repeatedly tested by the number of writers and their associated texts. After about five attempts, the figure of ten scribes

31 Tikhonov, *Sprachen* (2022).

32 Graves et al., *Connectionist* (2006).

came as a plausible result, in which the methods of machine and philological classification were compared.

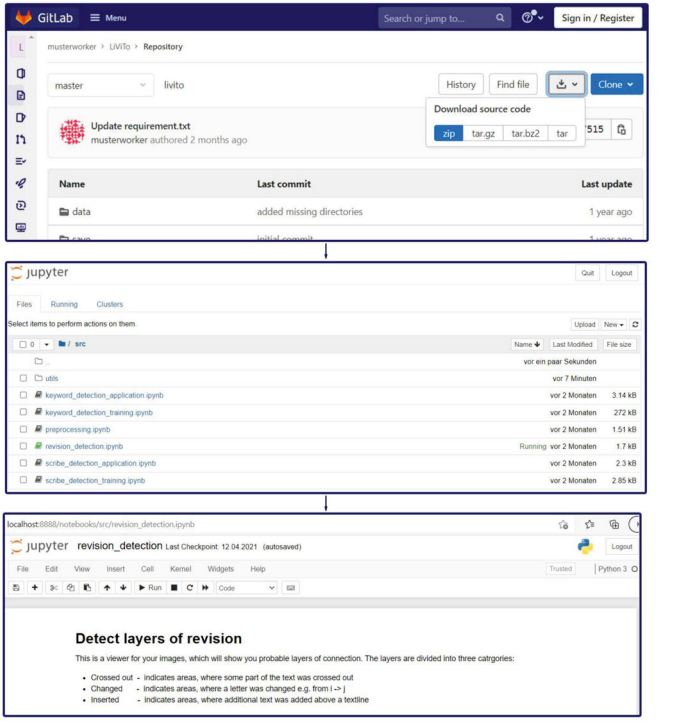
G. Status of data visualization

The stylometric analysis of the manuscripts uses a rather limited set of linguistic features, deliberately chosen by the researcher, to calculate distance matrices between the documents. They can be presented either in tables or in various types of graphs; and graphs constitute a virtually indispensable tool for the identification of clusters in the data. Machine-based optical pattern recognition, moreover, is based on an analysis of no less than 128 different visual handwriting features. Here, the only appropriate way of depicting the results is to visualize the distances between various clusters of similar manuscripts.

We ensure visualization in LiViTo as a research assistance tool by using open-source software Jupyter Notebook. The web-based interactive environment as part of the Jupyter Project makes LiViTo available as web browser-based application.[33] All three functions of LiViTo—localization of revisions, keyword spotting, and clustering according to visual characteristics—can be started and used as three separate applications in the Jupyter Notebook. The users have to be generally open to programming languages, but they do not have to be able to code. A detailed ReadMe document contains commands and preparatory installation steps that must be conducted by the user. Once these pre-settings are done, the functions of the program run with very little effort. All three LiViTo functions have a maximum of 10 short steps that lead to a result or partial result and are described in the ReadMe document. The user can extract and download her/his results from the Jupyter Notebook with just a few clicks (cf. https://gitlab.com/musterworker/livito).

33 Adam Rule, Amanda Birmingham, Cristal Zuniga, Ilkay Altintas, Shih-Cheng Huang, Rob Knight, Niema Moshiri, et al., "Ten Simple Rules for Writing and Sharing Computational Analyses in Jupyter Notebooks", in: *PLOS Computational Biology* 15/ 7, (San Francisco: PLOS, 2019).

Fig. 5: LiViTo in the jupyter notebook's interface in a browser on Windows 10; top down: downloading LiViTo repository for the intallation, LiViTo's functions.



Discussion

A. Digitalization:[34] Increase in efficiency and change of research perspective

Normally, working with old manuscripts implies going to an archive or library, putting on white gloves, and leafing carefully through pages, taking notes or transliterating the content. This method applies to philologists, historians, theologians and many other scholars who work with such documents. Some institutions

34 For a terminological differentiation between digitization and digitalization see Mary Anne M. Gobble, "Digitalization, Digitization, and Innovation", in: *Research-Technology Management* 61/4, (Virginia: Industrial Research Institute, 2018), 56–59.

offer scanning services, or researchers are allowed to take photos of the manuscripts for free or for a fee so that they can work with scans or photos without time restriction. All the traditional methods are rather impractical in cases of research on hundreds or thousands of pages.

Digitalization of manuscripts brings benefits to both sides—to computer science and to the humanities. Digitalization forms the ground truth for machine learning and for the definition and collection of quantitative linguistic features, and it allows a very detailed examination of the documents which is essential for developing the revision detector function of LiViTo, so that it works for every examination case without uploading transliterated documents. The function has been successfully tested with German, Czech, French, Hebrew and Latin. In addition, digitalization was helpful for the linguistic part of the project, because it made legible marginalia that were no longer discernible to the naked eye. This revealed new facts about the manuscripts. Several handwritten copies or versions of these manuscripts were localized in the Czech Republic, and the complete history of the manuscripts could be traced. With only classical Close Reading methods of the material or photographed manuscripts, none of these results could have been achieved. Working with digitized documents also enabled simultaneous and efficient co-working on the same pages or parts of the manuscripts.

Both sides of the project made different visual segmentations of the documents. Line course detection and line comparison became possible and further development of the keyword spotting function could be witnessed, in which not only whole words but also letters in the beginning, end, or in the middle of the word could be searched. This enables queries for roots, stems, prefixes or derivative affixes in terms of morphology. It also allows the identification of certain registers that are characterized by specific endings. In addition, a layout analysis was carried out at some points in the handwritten books. Subsequently inserted lines or entire passages were recognized. The texts could also be separated according to different principles (e.g. grammatical person—first or third),[35] whereby the hybrid authorship or collective vs. individual genesis of the manuscripts was proved.

B. Methodological controversies

Regarding our philological and historical scientific community, there were no problems presenting the project at colloquia and conferences. The absolute majority of colleagues reacted with great interest and eagerly awaited the results. Only one specific misunderstanding, which concerns the combination of quantitative and qualitative methods, came up several times and had to be clarified. Some colleagues conceived of the project goal as a complete switch to quantitative methods and

35 Tikhonov, *Sprachen*, 109.

optimization of research tools and procedures. On the contrary, the quantitative approach without the qualitative one would only yield partial results (and vice versa)—the combination of both was absolutely instrumental. The interaction of computational and linguistic methods was decisive for the success of the project.

To demonstrate this with a concrete example, the 3D graphical representation of the clustering in scribe identification is the result of at least three large comparison tests over approximately 12 months. In the beginning, linguistic features were combined with the visual features automatically recognized by AI methods. Next, the results of independent computational and linguistic analyses were compared. After each comparison, the analytical criteria were improved in accordance with the partner method. The 3D clustering then became a manageable result for the analyses.

However, a profound interpretation of this clustering is not possible without a deeper philological analysis. Quantitative visual and linguistic features were used on both sides in order to achieve a common quantitative result. This result then has to be translated into qualitative findings on both sides. In the literal sense of the word, we must zoom in on each individual point of the cluster diagram in the application and take into account the non-visual and quantitative-linguistic features in order to ultimately state concretely how many people wrote the documents and who these people were. So classical qualitative methods are by no means irrelevant—they just need to be combined with quantitative approaches.

C. Details versus abstraction

LiViTo provides both options: details and abstractions. The search results can be presented as a general overview or in detail. Depending on the research question, there are different relevant types of results—small but meaningful details or general overviews of large amounts of analyzed research data. As for the question of quantity and quality, we do not argue for an 'either/or' principle, but rather for a balanced combination. Both approaches benefit from each other. The task of the researchers is to use the right method at the right point of investigation.

Often it cannot be defined from the beginning that the research question will only be answered qualitatively or quantitatively, but there can be different scenarios. In the case at hand, qualitative preliminary examinations were carried out both in the computational and linguistics parts of the project. We first went into detail, that is, recognized prominent linguistic features and the regularities or irregularities in their occurrence; at the same time, we selected representative manuscript pages for first visual tests. In a stepwise process, we enlarged the amount of research data to be handled until we were able to take into account all the necessary features and all of the document pages. As soon as we achieved a first result for the full range of data, we checked whether it was realistic or it contained obvious errors both at a macro-

and micro-level (overview vs. detail). When details led to corrections, they had to be scaled up again in order to check for improvements at the more abstract level.

D. Towards a prototype DH laboratory

We are certainly no laboratory in the sense of a permanent institution. To us, a laboratory involves a larger set of researchers from the institutions to which the partners belong (HU, Fraunhofer IPK and MusterFabrik Ltd.), who contribute expertise from a wide range of fields. But we are certainly a team of scientists from different disciplines (including computer science and linguistics), who jointly and regularly conduct research on a common question, by using a mix of methods from their respective fields, in order to produce a joint result.

The most important phase in this common endeavour is the integration of research methods on the way to the concrete answer to a research question. Both sides complement each other with their competence in theoretical and practical areas; the result, however, is a common analysis rather than a confrontation of (computational vs. humanities') standpoints. In our experience, the integration phase has been the most time-consuming and rewarding part of our work, more intense than the actual formulation of results. It seems that such a level of intensity of exchange distinguishes a laboratory from a more loosely defined research group. In this sense, the project can be seen as a prototype DH laboratory. Based on this and several similar smaller-scale projects in the humanities and social sciences, HU Berlin has recently launched a long-term centre for 'Digitality and digital methods at Central Campus', headed by Roland Meyer and Torsten Hiltmann, Professor of Digital History.

Major outcomes and prospects for future DH research

We consider the major outcomes of our project to be

- (i) a better understanding of the respective contributions of machine-learning and linguistic/stylometric approaches to the task of detecting scribes and authors of historical manuscripts;
- (ii) an open-source software package which may assist researchers in detecting authors and scribes on larger sets of unknown historical documents;
- (iii) the concrete analysis of document origin and transmission for the 18th c. Czech autobiographies from the *Archiv im Böhmischen Dorf*, Berlin; and
- (iv) implications of this analysis for the history of Czech-German language and cultural contact in Berlin, and for the history of the Brethren.

If we focus on the more general DH-related aspects (i)–(ii) here, the obvious future prospect is the application of the mixed-methods approach of this project and its software prototype to other cases of author/scribe detection in other languages and historical periods. Already, within the small field of Slavic philology, many instances of unclear or disputed document origins come to mind, for example the older Church Slavonic witnesses that exist only in numerous partially overlapping later versions,[36] or texts of doubtful authenticity such as the Czech *Rukopisy královédvorský a zelenohorský*. Since the LiViTo tool is basically language-independent and requires only a very limited amount of training data, these possibilities will certainly be explored.

In the case of the Rixdorf Czech manuscripts, we have started to apply these techniques to the large set of sermons with promising first results. While most of them are obviously translated from German, their origin and transmission is interesting for the history of the Brethren mission; and there are many issues in historical linguistics which can be fruitfully approached on the basis of such a translation corpus. At present, we are exploring the translations of the Brethren sermons into other early modern languages (even rather exotic targets of missionaries) and their potential for creating a historical parallel corpus.

Sustainable access to digitized sources and to research data in general is becoming increasingly important. A significant branch of DH focuses on document preservation and digital archiving. In order to provide sustained availability of the digitized sources developed in the project, we intend to explore integration into the Laudatio repository[37] after consultation with the Brethren Archive.

During the last few years, character recognition technology (OCR) and handwritten text recognition (HTR) for manuscripts has witnessed most impressive developments that have opened up possibilities unheard of at the outset. In LiViTo, this has already led to effective string comparison even for untranscribed texts with a truly manageable training effort, enabling searches in such documents. Projects such as Transkribus[38] or eScriptorium[39] continue to foster progress in domains like line detection and OCR in historical manuscripts. It will certainly be rewarding to integrate components of these projects into the workflow of LiViTo in order to further improve our scribe detection.

---

36 For a fundamental non-DH treatment of the Slovo o zakone i blagodati see Giorgio Ziffer, "Jazyk i stil' slova 'O zakone i blagodati'", in: *Učënye zapiski Kazanskogo universiteta* 155 (5) (Kazan': Kazanskij (Privolzhskij) federal'nyj universitet, 2013), 7–16.

37 „LAUDATIO – Long-term Access and Usage of Deeply Annotated Information", Humboldt University Berlin, accessed February 2, 2023, https://www.laudatio-repository.org/.

38 "Transkribus – Where AI meets historical documents", READ-COOP, accessed February 2, 2023, https://readcoop.eu/transkribus/.

39 "eScriptorium", GitLab, accessed February 2, 2023, https://gitlab.inria.fr/scripta/escriptorium.

Independently, the linguistic side of the coin has witnessed considerable progress in the application of stylometry, which we intend to reflect in further research in the historical domain. Altogether, it seems that the main idea of the project—to combine pattern recognition for scribe detection and linguistic/stylometric analysis for authorship in order to uncover document origin and transmission for historical manuscripts—is as interesting and topical as ever. We hope that the integration of mixed methods achieved in the project together with the LiViTo tool will make a useful contribution to this area of research.

Bibliography

Burrows, John F. "Word-Patterns and Story-Shapes: The Statistical Analysis of Narrative Style". In: Literary and Linguistic Computing 2, 61–70. Oxford: Oxford University Press, 1987.

Burrows, John F. "'An Ocean Where Each Kind…': Statistical Analysis and Some Major Determinants of Literary Style". In: Computers and the Humanities 23, 309–321. New York/Heidelberg/AA Dordrecht: Springer, 1989.

Eder, Maciej. "Does Size Matter? Authorship Attribution, Small Samples, Big Problem". In: Digital Scholarship in the Humanities 30, 167–182. Oxford: Oxford University Press, 2010.

Eder, Maciej, Jan Rybicki, Mike Kestemont. "Stylometry with R: a package for computational text analysis". In: R Journal 8 (1), 107–121. Online-Open-Access-Publication, 2016. https://journal.r-project.org/archive/2016-1/eder-rybicki-kestemont.pdf.

Gobble, Mary Anne M.: "Digitalization, Digitization, and Innovation". In: Research-Technology Management 61/4, 56–59. Virginia: Industrial Research Institute, 2018.

Graves, Alex, Santiago Fernández, Faustino Gomez, Jürgen Schmidhuber, "Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks". In: Proceedings of the 23rd International Conference on Machine Learning, 369–376. Pittsburgh: Carnegie Mellon University, 2006.

Hauger, J. Scott. Reading Machines for the Blind: A Study of Federally Supported Technology Development and Innovation (Dissertation). Blacksburg: Virginia Polytechnic Institute and State University, 1995.

Hope, Jonathan. The authorship of Shakespeare's plays. A socio-linguistic study. Cambridge: Cambridge University Press, 1994.

Kahle, Philip, Sebastian Colutto, Günter Hackl, Günter Mühlberger. "Transkribus – A Service Platform for Transcription, Recognition and Retrieval of Historical Documents". In: 14th IAPR International Conference on Document Analysis and Recognition (ICDAR), 19–24. Kyoto: IEEE, 2017.

Kay, Anthony. "Tesseract: An Open-Source Optical Character Recognition Engine". In: Linux Journal. Online-Open-Access-Publication, 2007. https://www.linuxjournal.com/article/9676.

Kleinwächter, Livia. "The Literary Manuscript: A Challenge for Philological Knowledge Production". In: Philology in the Making Vol. 1, edited by Pál Kelemen and Nicolas Pethes, 109–128. Bielefeld: transcript Verlag, 2019.

Koppel, Moshe, Jonathan Schler, Shlomo Argamon. "Computational Methods in Authorship Attribution". In: JASIST 60, edited by Steven Sawyer, 9–26. Hoboken: Wiley-Blackwell, 2009.

Mettele, Gisela. Weltbürgertum oder Gottesreich: die Herrnhuter Brüdergemeine als globale Gemeinschaft 1727 – 1857. Göttingen: Vandenhoeck & Ruprecht, 2009.

Motel, Manfred. Das böhmische Dorf in Berlin: die Geschichte eines Phänomens. Berlin: Darge Verlag, 1983.

Müller, Klaus, Aleksej Tikhonov, Roland Meyer. „Livito: Linguistic and Visual Features Tool for Assisted Analysis of Historic Manuscripts". In: Proceedings of the Twelfth Language Resources and Evaluation Conference, 885–890. Marseille: European Language Resources Association, 2020.

Plecháč, Petr. "Relative contributions of Shakespeare and Fletcher in Henry VIII: An analysis based on most frequent words and most frequent rhythmic patterns". In: Digital Scholarship in the Humanities 36, 430–438. Oxford: Oxford University Press, 2021.

Rule, Adam, Amanda Birmingham, Cristal Zuniga, Ilkay Altintas, Shih-Cheng Huang, Rob Knight, Niema Moshiri, et al. "Ten Simple Rules for Writing and Sharing Computational Analyses in Jupyter Notebooks". In: PLOS Computational Biology 15/ 7. San Francisco: PLOS, 2019. https://doi.org/10.1371/journal.pcbi.1007007.

Seibt, Angelika. Unterschriften und Testamente – Praxis der forensischen Schriftuntersuchung. München: Beck, 2008.

Sterik, Edita. Die böhmischen Exulanten in Berlin. Herrnhut: Herrnhuter Verlag, 2016.

Štěříková, Edita. Běh života českých emigrantů v Berlíně v 18. století. Praha: Kalich, 1999.

Tikhonov, Aleksej. Autorenidentifikation und linguistische Merkmale der Rixdorfer Handschriften: Eine Untersuchung anhand von Manuskripten aus dem 18./19. Jahrhundert (Dissertation), Berlin: Humboldt-Universität zu Berlin, 2020.

Tikhonov, Aleksej. Sprachen der Exilgemeinde in Rixdorf (Berlin): Autorenidentifikation und linguistische Merkmale anhand von tschechischen Manuskripten aus dem 18./19. Jahrhundert. Heidelberg: Winter Verlag, 2022.

Tikhonov, Aleksej and Klaus Müller. „Livito: A software tool to assess linguistic and visual features of handwritten texts". In: Qurator – Conference on Digital Cura-

tion Technologies 2020, edited by Adrian Paschke, Clemens Neudecker, Georg Rehm, Jamal Al Qundus, Lydia Pintscher. Berlin: Online-Open-Access-Publication, 2020. https://ceur-ws.org/Vol-2535/paper_8.pdf.

Tikhonov, Aleksej and Klaus Müller. "Scribe versus authorship attribution and clustering in historic Czech manuscripts: a case study with visual and linguistic features". In: Digital Scholarship in the Humanities 37, 254–263. Oxford: Oxford University Press, 2022.

Ziffer, Giorgio. "Jazyk i stil' slova "O zakone i blagodati"". In: Učënye zapiski Kazanskogo universiteta 155 (5), 7–16. Kazan': Kazanskij (Privolzhskij) federal'nyj universitet, 2013.

Zyl, Maryka van and Yolande Botha. "Stylometry and Characterisation in The Big Bang Theory". In: Literator 37/ 2, 1–11. Cape Town: Aosis Publishing, 2016.

#UNCERTAINTY

The OED defines #Uncertainty among others as the "quality of being uncertain in respect of duration, continuance, occurrence, etc." and as "liability to chance or accident". Also included in the definition is "the quality of being indeterminate as to magnitude or value; the amount of variation in a numerical result that is consistent with observation". The definitions relate, among others, to vagueness in the sense of not knowing and, in economics, to risk by a lack of predictability. *

Thus, while uncertainty in general is a core phenomenon in all fields of research, its meaning is especially ambivalent in digital humanities when the quality of not knowing and a quantifiable category of uncertainty in a statistical sense, meet in communication.

Most mixed methods projects address this at one or the other level, sometimes discerning #Uncertainty from vagueness especially in the context of linguistics. They underline #Uncertainty as intrinsic or inherent to data; its transformation and interpretation (HerCoRe, Digital Plato; cf. also DhiMu, ArchiMediaL) or address having varying levels or qualities of #Uncertainty either as a result of #Quantification (DhiMu) or as an outcome of #Machine Learning (ArchiMediaL).

The concept of #Uncertainty denotes significant epistemic challenges, but despite its ambiguous role in digital humanities and mixed methods processes, it does not seem to affect communication between the parties involved.

* "Uncertainty, n.". in: *Oxford English Dictionary* (OED), first published 1921; most recently modified version published online March 2022, https://www.oed.com/ [accessed: 25.10.2022].

Title: HerCoRe – Hemeneutic and Computer-based Analysis of Reliability, Consistency and Vagueness in Historical Texts
Team: Cristina Vertan (digital humanities, computational Linguistics)-PI, Alptug Güney (Turcology), Walther v. Hahn (German Computational Linguistics, Digital Humanities), Ioana Costa – University of Bucharest (Latin and Romanian Linguistics), Rafael Quiros Marin (GIS-Application and Web Frontend), Miguel Pedegrosa (Web Frontend), Anja Zimmer (mathematical foundation fuzzy logic) Corpus, Yavuz Köse –University of Vienna (Turkology)
Mutlilingual corpus: Two most important works of Dimitrie Cantemir (humanist of the 18th century) "Description of Moldavia" and "History of the Growth and Decay of the Ottoman Empire" in Latin, German (original translations from the 18th century) and Romanian
Field of Study: Turcology, Balkan /Romanian History
Institution: University of Hamburg
Methods: data modelling (Graph, RDF-Triple, Objects), natural language processing, hermeneutics, source manual analysis, ontology development, software development, annotation, fuzzy reasoning
Tools: Protégé (Ontology), OrientDB (Database) , WebLicht (Language Technology), CDG-Parser (Language Technology), HistAnno (in project developed Annotation tool) (Annotation), Oxygen (XML Edit). Eclipse (Software development)
Technology: OWL, Java EE, Javascript , Document and Graph Databases, Leaflet, SPARQL

Encoding, Processing and Interpreting Vagueness and Uncertainty in Historical Texts – A Pilot Study Based on Multilingual 18th Century Texts of Dimitrie Cantemir (HerCoRe)

Cristina Vertan

Abstract *Qualitative and Quantitative Analysis of historical documents by means of computational methods can lead to a better understanding of the particular texts but also of the époque to which they relate. This can be however achieved only if 1.) the digital methods encode, embed and allow visualisation of the various manifestations of vagueness and ambiguity present at all levels in text (metadata, edition, discourse, lexical etc.), 2.) the knowledge-base is built in close cooperation with specialists in the respective time period and 3.) the final interpretation is left to the user. In this contribution we will present an innovative approach for dealing with the vagueness of natural language and show how Mixed Methods can help in detecting reliability and consistency of historical documents from the 18th century. We use a multilingual corpus consisting mainly of texts in Latin, German and Romanian and apply methods from fuzzy logic, graph theory, knowledge acquisition and natural language processing.*

Introduction

Over the last decades, digital processing of historical texts (in fields such as Digital History, Historical Linguistics or Digital Culture Heritage) has focused primarily on digitization and adaptation of computer linguistic tools for (nearly) extinct languages (e.g. Classical Greek, Latin, or Coptic), and on old language variants (e.g. Middle High German or Old French). Massive digitization campaigns not only allowed preservation of historical documents but also increased significantly the distribution of these texts to a broad spectrum of researchers. However, only in recent years has a movement occurred from 'digital reading' (search through digital catalogues and display of texts with progressive zoom facilities) to 'digital analysis' of the texts. However, it goes rarely beyond keyword search and quantitative (i.e. statistical) measurements and lacks interpretation/contextualization/specification of

individual words and terms. This may lead to inaccurate/wrong interpretations because, for example, words have different meanings according to the context and simple statistics about co-occurrences of two terms do not automatically imply causality between them ('Romans were eating a lot of bread' and 'Romans died younger' are two true assertions but they do not automatically imply that 'Romans died younger because they were eating a lot of bread'. For the latter utterance one may need more detailed analysis beyond statistical measurements).

Approximate dates, unclear places and less documented existence of persons are just some examples of types of vagueness and uncertainty encountered in such documents. Ignoring this characteristic of historical documents not only may lead to an incomplete digital recording, but also has important consequences on the interpretation process whether it is done by man or machine. The project HerCoRe (Hermeneutic and Computer-based Analysis of Reliability Consistency and Vagueness in Historical Texts) has investigated, in a concrete multilingual corpus, how vagueness and uncertainty can be recorded digitally and used for hermeneutic interpretation. Most of the current approaches tend to reduce expressions like 'towards the beginning of the 18th century' to crisp values like 1700, or in best case an interval 1700–1720. An event happening in 1721 is thus no longer considered early of the 18th century. Vague expressions such as 'it is probable that the event occurred' are reduced to 'the event occurred'. Thus important information is omitted and the analysis lacks accuracy whether it is performed by computers or humans (hermeneutic).

Methods from semantic web and natural language processing have the potential to enrich these digital historical recordings with intelligent functionalities, so that new insights in the materials are gained or hitherto hidden connections between events or persons are revealed and can be visualized.

The basis of all these new functionalities is in many cases a deep annotation of the contents. Methods of machine learning by using large training resources are difficult to be applied because these training resources are missing. Additionally the embedding of vagueness and uncertainty in the analysis (throughout a genuine method in hermeneutic approach) is often neglected Following a brief analysis of the state of the art in section 2, we proceed to explore the reliability and consistency of two works of Dimitrie Cantemir, a humanist of the early 18th century and expert (among others) in the culture and history of the Ottoman Empire, by using advanced computer science tools such as ontologies and inferences. We show how uncertain and vague information, so often neglected by state-of-the-art Digital Humanities approaches, can be modelled and included in the computer-aided analysis of texts.

Overview of State-of-the-Art Annotation of Vagueness and Uncertainty in Digital Humanities (DH)

Language is one of the most powerful tools people use to express themselves and their cultural as well as social environment. Words have meanings, follow certain syntactic rules and many are just labels for concepts. In real world, it is quite common that information cannot be classified as 100% true or false. However, all these features we associate with a word are part of our background knowledge. Computers represent words as a sequence of '1' and '0'. All background information (semantic, syntactic, or conceptual) has to be supplied through annotations.

Many annotation systems implemented up to now tend to ignore an intrinsic character of natural language which is of great importance for the analysis of historical documents: the vagueness of almost every utterance.[1] Two parameters, vagueness and uncertainty, are mostly neglected in annotations of historical documents. Studies on natural language have shown that even in technical domains (e.g. manuals for equipment) language expressions are vague. Texts dealing with past events and written at a time when access to information was restricted are full of utterances expressing doubt or remaining vague about events or places. Additionally, variations in spelling, use of oral tradition instead of documented sources lead to uncertainty especially about named places (e.g. 'not far away from the Danube'), persons ('known as the son of…') and dates (e.g. 'after the last call of the muezzin').

TEI (Text Encoding Initiative)[2] is the state-of-the-art standard for annotation of historical documents. It offers three possibilities for recording (i.e. annotating) vagueness.

1. Using the <note> element the user can write unstructured text, mentioning the degree and scope of the identified vague aspect.
2. The <certainty> element: it offers the possibility of structuring the information about vagueness in the following ways. The <certainty> element can refer to the name of the annotation tag considered uncertain (e.g. a person or a place name), the position in text where the annotation tag starts, or a value of an attribute contained in the annotation tag. Through the attribute @degree it is possible to refine the level of certainty. The <certainty> element can refer to one or more annotation elements through XPath expressions.
3. The <precision> element, which can be applied for any numerical value (a date, or a measure), indicates the numerical accuracy associated with some aspects

1 Walther von Hahn, "Vagheit bei der Verwendung von Fachsprachen", in: Lothar Hoffmann, Hartwig Kalverkämper and Herbert Ernst Wiegand (eds.): *Fachsprachen. Ein Internationales Handbuch zur Fachsprachenforschung und Terminologiewissenschaft* (Berlin 1998), pp. 383 – 390.

2 www.tei.org.

of text markup. If a standard value is precise and known, the user can express it with the element <precision> and the attribute @stdDev, which represents its standard deviation.

Additional TEI offers the possibility of indicating the responsibility for the whole content or partial annotators. In this way the user can specify if the vagueness is due to the author, the quoted source or the editor. Such detailed information is usually done manually.[3]

However, if we want to annotate vague expressions or other levels of uncertainty, we can face several drawbacks of the TEI approach:

1. Overlapping annotations concerning vagueness are possible only as stand-off annotation. However, stand-off annotation in TEI is extremely complicated.
2. There are different levels of vagueness introduced by the author by the referred source, dating etc. Not all of these sources of vagueness can be specified with the <response> tag which can be attached just to individuals.
3. The <precision> element can be specified just for numerical values. An expression 'some kilometres south from the city' introduces a non-numerical vague coordinate. When we speak about historical documents, geo-location of the place is not always possible.
4. There is no reasoner (computer programme for realizing logical connections and inferences between assertions) which can be applied to the TEI annotation.

The HerCoRe project has developed an annotation-tool which is able to handle drawbacks 1 and 2.

HerCoRe Case Study for Mixing Methods from Humanities and Computer Sciences

As mentioned above, digital processing of historical documents is particularly challenging with regard to vagueness and uncertainty. Approximate dates, unclear places and less documented existence of persons are just some examples of the types of vagueness and uncertainty encountered in such documents. Ignoring these characteristics of historical documents may lead not only to an incomplete digital recording but also to important consequences for the interpretation process whether it is done by man or machine.

3 More details and examples of using these TEI-elements can be found on https://web.uvic.ca/lancenrd/martin/guidelines/tei_CE.html.

The project is investigating Latin manuscripts and their translations in German and Romanian of two most relevant works on the Balkan history, written in the middle of 18th century by Dimitrie Cantemir. Section *Rationale* introduces the author, his relevance for Ottoman and Balkan studies and describes the still unsolved research issues. In section *Corpus*, we provide details about the corpus, the transmission and the selection of texts for the HerCoRe project. Finally, in section *Mixed-Method Investigation Approach* we present the mixed-method approach including a pure humanistic (hermeneutic) approach as well as modern technologies from computer science.

Rationale

For historians, research on the Ottoman Empire is challenging given the time interval (more than 700 years) and the number of languages and cultures involved. Many documents in Turkey or the Balkans have become accessible physically only in recent years; still they cannot be fully analysed due to language or script barriers. Only within the Balkans (as a former part of the Ottoman Empire) do we encounter four alphabets (Latin, Cyrillic, Arabic and Greek) and several national languages in addition to the ones considered official at the time of Ottoman Empire (Greek and Slavonic in the churches, Ottoman Turkish and sometimes also Latin). For many centuries, information about the Ottoman Empire was sparse and limited to a few diplomats and travellers who reported it in a fragmentary way. In this context the works of Dimitrie Cantemir represent an essential milestone in the reception of the Ottoman society in Western Europe. Born in Romania in 1673 as the son of a Moldavian ruler (Wojevode), Cantemir was sent to Istanbul as guarantee for the loyalty of his father. He continued his education (which he begun in Moldavia with Greek monks) in Istanbul and became one of the last universal scientists of the Age of Enlightenment. D. Cantemir published works in history, geography, philosophy and music, being the first one to record Ottoman music on paper. Of particular relevance were two of his works, written at the request of the Royal Academy of Sciences in Berlin, of which he became a member: 'Descriptio Moldaviae' (The History of Moldavia, DM) and 'The Rise and Decay of the Ottoman Empire' (HO). These works were written in Russia, at the court of the Tsar Peter the Great, where he received asylum after a short period of ruling over Moldavia. This detail is important as it may have implications about the two mentioned works: it is not clear if historical mistakes in his writings were due to lack of sources or were done deliberately in order to comply with the ideology and the religion of his new protector, the tsar.

The two works mentioned above reached London after his death, being brought by his son as a diplomat. They were translated by Georges Tindal[4] immediately. The

4 https://en.wikipedia.org/wiki/George_Tindall.

originals were lost shortly afterwards. Tindal's translations were used for French, German and, later in the 19th century, Romanian versions. All these translations represented for more than one century the most comprehensive information about the Ottoman Empire. Even later historians at the end of 19th and beginning of 20th century, who claim to quote Cantemir, quote in reality one of these translations, especially the English and the German one. In the 1920s, several researchers raised doubts about the accuracy of Cantemir claims. Again, they compared translations with historical facts described, meanwhile, by other researchers. Only in the second half of the 20th century were the Latin manuscripts of Cantemir's works discovered (three manuscripts for Descriptio Moldaviae and one for History of the Ottoman Empire). Even the shallow comparison between these manuscripts and the translations revealed that Tindal had made serious deviations from the original, for example, paragraphs written in Arabic were omitted and some facts or descriptions of persons or traditions were changed. We will describe in detail these differences in section *Corpus*.

For researchers in Ottoman studies, Cantemir's writings are of great importance, yet a thorough comparison of existing manuscripts and translations is still lacking. This aspect could not have been realized until today as the number of scripts and languages involved as well as the size of the material are difficult to be handled by humans. Additionally, recent digitalization campaigns in Turkey give access for the first time to documents from the 18th century, which may support Cantemir's claims.

We focus our research on three directions:

1) Reliability: Questions to be investigated here are: are the quotations made by Cantemir himself grounded in authenticity? Is there concordance between his degree of trust in these sources and the current knowledge about them (e.g. is there any evidence that a person which Cantemir claims to have spoken to, really lived in that time?). In the translations, too, the insertions of the editors or translators have to be investigated.
2) Consistency: Does Cantemir keep a constant opinion about persons, events, facts across the same work (e.g. in the main text and his own annotations)? The two main works of the corpus have a common pool of persons, events and places; thus we want to investigate also if the author keeps to his opinion or describe things according to the target public and possible political context.
3) Vagueness: Cantemir's works are full of vague expressions. He states that even 'I do not dare to decide what the truth is about this matter, given the high darkness of this story'. The corpus analysis tried to emphasize these expressions and analyse if they were used on purpose (political or tactical reasons) or just as a stylistic tool. Also, we have a look at the translations in order to see to what extent they preserve the degree of vagueness stated by Cantemir.

The project HerCoRe, therefore, addresses a real important research problem for Ottoman studies. At the same time, the complexity of the corpus and of the user requests (i.e. the historians' demand for a digital tool) calls for a different approach to the current state-of-art in digital humanities. We will discuss it in detail in section *Mixed-Method Investigation Approach*.

Corpus

As mentioned in section *Rationale*, our research focuses on the two books of Dimitrie Cantemir covering the history of the Ottoman Empire (HO) and the description of his country Moldavia (DM). From the numerous translations and editions, we chose:

For HO: the edition of the Latin manuscript found in Harvard University, the translation in Romanian of this edition, and the historical translation in German from 1771.[5]

For DM: the edition collating the three existent Latin manuscripts from St. Petersburg and Odessa,[6] the translation in Romanian of this edition[7] and the historical translation in German from 1745.[8]

The choice had a scientific and a pragmatic reason. From the scientific point of view, we used in both cases the most recently available editions which collated all historical manuscripts known until now. The Latin editions and the Romanian translations were realized by well-known specialists, one of them being part of the HerCoRe team. The German historical translation is relevant for two reasons. Firstly, it was done after the English translation of Tindal preserved most of his errors but corrected some of them. Secondly, it was one of the most used in the 18th and 19th centuries, especially for further translations. In this respect, it offers a different narrative to those of the recent editions and translations.

From a pragmatic point of view, we chose those documents that were also available in machine readable versions (with one exception[9]). Tindal's translation is for the moment available only in PDF format. Given the number of foreign words in the text and the relatively low PDF-quality of existent files, any PDF-to-text conversion would have failed. Given the time span of the project, digitization of further documents was not considered.

---

5 Dimitrie Cantemir, *Beschreibung der Moldau*, Faksimiledruck der Originalausgabe von 1771 (Frankfurt und Leipzig: 1771).

6 Florentina Nicolae and Ioana Costa (eds.), *Descrierea stării Moldaviei: în vechime şi azi* (Bucureşti: Academia Română-Fundaţia Naţională pentru Ştiinţă şi Artă, 2017), 2019.

7 Nicolae and Costa, *Descrierea*.

8 Dimitrie Cantemir, *Geschichte des osmanischen Reichs nach seinem Anwachse und Abnehmen* (Hamburg: Herold, 1745).

9 Cantemir, *Beschreibung*.

The HerCoRe Corpus contains three types of texts, each of them with its own particularities:

- 1. Editions of Latin manuscripts[10]
  These documents include not only the original text but also markup and footnotes of the editor. The markup (e.g. different bracket types) has to be removed from the basic text and reinserted as annotations. They represent a challenge for any tool processing natural language. The main language of the manuscripts is Latin. However, Cantemir reproduces names of persons, places and quotations in original Ottoman Turkish written with Arabic characters and also provides a Latin transliteration of the same. Often the manuscript contains a translation of these paragraphs into Latin. Also, Greek quotations are encountered. Romanian words are written (in contrast with the tradition in the 18th century) with the Latin alphabet. As far as it is known, it is the first attempt of using Latin instead of Cyrillic alphabet for this purpose. The transcriptions are however not standardized and, as a consequence, a name has several transcriptions. For example, the name of the Moldavian capital (Iași) is encountered in 32 transcription forms (Iassi, Jassy etc.).
- 2. Modern translations of Latin manuscripts into Romanian[11]
  In these documents the translators inserted footnotes with explanations. They retained the transcriptions of Romanian names as in the Latin manuscript and at places offered explanations in footnotes. Turkish and Greek phrases were preserved in the original alphabets and transcriptions. The translator kept Cantemir's transcription style even though it did not correspond to the modern Romanian language. The translation is, however, adapted to suit the modern reader and uses to a great extent current Romania grammar and lexicon.
- 3. Historical translations into German[12]
  These translations mainly came after the historical English translation by Tindal and (as the editors mention in the foreword) with an additional consultation in case of DHO of the French translation (which however had as basis this English version as well). The Ottoman Turkish paragraphs were omitted completely (Arabic original and Latin transcription). If Cantemir offered a Latin translation of the paragraphs then they consequently appeared in German. In all other cases the Turkish quotations were simply omitted. Romanian and Turkish names and

10 Octavian Gordon, Florentina Nicolae, Monica Vasileanu and Ioana Costa (eds.), *Cantemir Dimitrie, Istoria măririi şi decăderii Curţii othmane*, 2 volumes (Bucureşti, Academia Română-Fundaţia Naţională pentru Ştiinţă şi Artă, 2015), Nicolae and Costa, *Descrierea*.

11 Costa et al., *Istoria*; Nicolae and Costa, *Descrierea*.

12 Cantemir, *Geschichte*; Cantemir, *Beschreibung*.

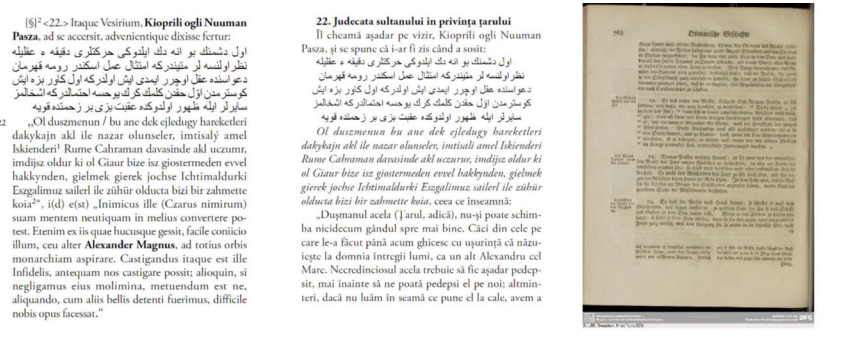
denominations of occupations were modified and adapted to German phonetics. For example, 'Kioprili ogli Nuuman Pasza' in the Latin original was preserved in the Romanian modern translation (although the modern form of Pasza would be in Romanian Paşa), but changed to 'Kjüprili Ogli Numan Pascha' in the German version.

These variations make practically impossible a consistent character-based search over the entire corpus.

The German translation also includes footnotes by the editor, marked with (V). Both German translations were printed with black-letter fonts (Fraktur). This is an additional challenge both for an untrained person and for the preparation of machine readable versions.

The example illustrated above can be retrieved in Figure 1.

Fig. 1: Same paragraph in Latin manuscript and translations. From left to right: Latin edition,[13] Romanian edition,[14] German edition.[15]



{§}[2] <22.> Itaque Vesirium, **Kioprili ogli Nuuman Pasza**, ad se accersit, advenientique dixisse fertur:

اول دشمنك بو انه دك ايلدوكى حركتلرى دقيقه ء عقليله نظر اولنسه لر متيندركه امتثال عمل اسكندر رومه قهرمان دعواسنده عقل اوجرر ايمدى ايش اولدركه اول كاور بزه ايش كوسترمدن اوّل حقدن كلمك كرك بوحسه احتمالدركه اشخالمز سايرلر ايله ظهور اولدوكده عقبت بزى بر زحمتده قويه

522 „Ol duszmenun / bu ane dek ejledugy hareketleri dakykajn akl ile nazar olunseler, imtisalý amel Iskienderi[1] Rume Cahraman davasinde akl uczumr, imdijsz oldur ki ol Giaur bize isz giostermeden evvel hakkynden, gielmek gierek jochse Ichtimaldurki Eszgalimuz sailerl ile zühür olducta bizi bir zahmette koia[2*], i(d) e(st) „Inimicus ille (Czarus nimirum) suam mentem neutiquam in melius convertere potest. Etenim ex iis quae hucusque gessit, facile coniicio illum, ceu alter **Alexander Magnus**, ad totius orbis monarchiam aspirare. Castigandus itaque est ille Infidelis, antequam nos castigare possit; aliaquin, si negligamus eius molimina, metuendum est ne, aliquando, cum aliis bellis detenti fuerimus, difficile nobis opus facessat."

22. Judecata sultanului în privinţa ţarului

Îl cheamă aşadar pe vizir, Kioprili ogli Nuuman Pasza, şi se spune că i-ar fi zis când a sosit:

اول دشمنك بو انه دك ايلدوكى حركتلرى دقيقه ء عقليله نظر اولنسه لر متيندركه امتثال عمل اسكندر رومه قهرمان دعواسنده عقل اوجرر ايمدى ايش اولدركه اول كاور بزه ايش كوسترمدن اوّل حقدن كلمك كرك بوحسه احتمالدركه اشخالمز سايرلر ايله ظهور اولدوكده عقبت بزى بر زحمتده قويه

Ol duszmenun bu ane dek ejledugy hareketleri dakykajn akl ile nazar olunseler, imtisalî amel Iskienderi Rume Cahraman davasinde akl uczurur, imdijsz oldur ki ol Giaur bize isz giostermeden evvel hakkynden, gielmek gierek jochse Ichtimaldurki Eszgalimuz sailerl ile zühür olducta bizi bir zahmette koia, ceea ce înseamnă:

„Duşmanul acela (Ţarul, adică), nu-şi poate schimba nicidecum gândul spre mai bine. Căci din cele pe care le-a făcut până acum ghicesc cu uşurinţă că năzuieşte la domnia întregii lumi, ca un alt Alexandru cel Mare. Necredincios ul acela trebuie să fie aşadar pedepsit, mai înainte să ne poată pedepsi el pe noi; altminteri, dacă nu luăm în seamă ce pune el la cale, avem a

These particularities are real challenges for the preparation of the machine readable corpus. In the following section, we will describe which steps had to be taken for an appropriate digital representation and the workflow plan we prepared in order to combine information technology and hermeneutic methods.

13 Nicolae and Costa, *Descrierea*.

14 Costa et al., *Istoria*.

15 Cantemir, *Geschichte*.

Mixed-Method Investigation Approach

The complexity of the research questions presented in section *Rationale* as well as the heterogeneity of the data as described in *Corpus* made an exclusively hermeneutic approach and a fully automatic processing exercise impossible. Thus a mixed method paradigm seemed to suit this research project. Our aim was to involve both computer scientists and researchers from relevant humanities fields (Ottoman history, German and Romanian linguistics, Romanian history and Latin studies) in all phases of the investigation. Our research built on three pillars: data modelling, data processing and data visualization as well as investigation. In this section, we will explain how a mixed approach was used for each of the three pillars.

Data Modelling

This step involves data acquisition followed by recording of data in an appropriate model. Additionally, we scrutinized here which additional knowledge was needed and how it could be provided in a digital form.

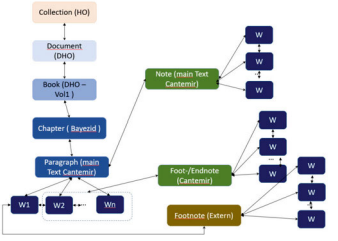
Four axioms guided our work:

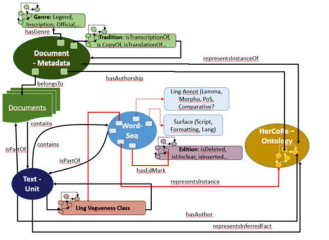
- A1. At any time, human researcher or automatic process should be able to identify the text source: main text, side- or end-annotation done by the author of the main text, editor's/translator's annotation.
- A2. Editorial markup in the Latin manuscript has to be preserved as long as it concerns insertions, removals or empty spaces in the manuscript.
- A3. Layout information (e.g. italic or bold formatting) underlies a meaning and has to be preserved.
- A4. Text normalization is done only in cases where rules can be applied without exceptions and they do not interfere in further analysis. This means we would not normalize across languages the transcription of proper names or specific occupations as these variations could be a source of misinterpretations and wrong translations. On the other side, we normalize the 18th century German diacritics style (e.g. $\overset{e}{u}$ -> ü, $\overset{e}{a}$->ä, ->s) as well as old writing rules for which non-ambiguous transformations are possible (e.g. bey->bei).

We use two models for our corpus. The first one, the Document Collection Model, defines the corpus structure. A Collection represents one work (HO or DM) including the Latin manuscript as well as the Romanian and German translations. Each collection contains several documents; in our case, one for each considered language. Substructures such as book, chapter and paragraph follow. The smallest unit in this structure is a word-unit (a string separated by empty spaces). Each word is recorded

as an object identified by a unique ID. These IDs can be used in order to define non-hierarchical data structures such as notes written either by the author or by the editor/translator. This model is presented in Figure 2.

Figure 2: Document collection model. *Figure 3: Annotation model.*





All additional information (linguistic, editorial, layout, knowledge, vagueness) is attached to groups of word-units and is used in the following annotation. In this way, we allow annotations over discontinuous group of word-units and overlapping of several annotation layers. The annotation model is presented in figure 3 and shows different annotation types (corresponding to different layers) defined on sets of word-units. In the system, a word-unit is an object with a unique identifier and several labels (normalized writing, original writing) and pointers to annotation objects. This model comes with great advantages: it allows different ways of display of the text including the changes or corrections made in the initial text (e.g. OCR corrections).

In the following section, we will discuss in detail the knowledge and vagueness annotations which are central aspects of our mixed-method research paradigm.

Levels of Vagueness and Uncertainty in HerCoRe —Challenges and Solutions

As mentioned in the previous sections, in the HerCoRe corpus we encounter different types of vagueness and uncertainty which need to be annotated in order to ensure a proper qualitative/hermeneutic analysis. The following classes of uncertainty /vagueness can be defined:

A. Linguistic

- *Linguistic uncertainty* refers to expressions of temporal or geographical information, for example, 'some years before', 'more or less 5 kilometres', 'two days far

from' or 'at the beginning of 13th century'. These expressions assume that a certain location exists or that a certain event has taken place but their geographical or temporal coordinates cannot be fixed. This type of uncertainty is witnessed mostly at the lexical level.

- *Linguistic vagueness* refers to those expressions that induce a subjective position of the narrative, for example, 'it was said', 'I am not sure', or 'it is supposed'. Such vagueness indicators may be lexical, or they may even be syntactic, for example, through the use of language-specific verb modalities or time: 'would have been', 'should have taken place'. This syntactic vagueness is challenging because it is difficult to say if it is just a matter of narrative style or a vagueness marker used intentionally.

B. Knowledge

- Proper names (geographical, persons). Uncertainty is often related to writing variants or political renaming of places over time which may lead to confusions (e.g. Izmir vs. Izmit, two different cities in the Ottoman Empire; Sultan Bayezid—there were two rulers with this name). Vagueness is connected with geographical areas with borders changing over time (e.g. Ottoman Empire, Principality of Moldavia) or regions for which the exact coordinates cannot be defined (e.g. Syrfia).
- Concepts. They are domains and in our case also country specific. From the linguistic point of view, the lexicalization of these concepts is not vague or uncertain. We can take for example, the word 'Vizier' (a political functionary in the Ottoman Empire and equivalent to a minister in modern times). Used in the plural, it denominates all such ministers, but when it is used in the singular it refers usually to the Great Vizier (the leader). Depending on the context, it can also point to a certain person named in a previous sentence (anaphora). Another example may be a 'place of pilgrimage' which has different definitions depending on the religion in question.

C. Sources (factual uncertainty)

- Uncertainty of sources refers to good documented sources (e.g. chronicles) for which certain parameters are not precise (e.g. author or publication date). However, the physical objects being referred to exist in many cases and they can be consulted.
- We classify as vague sources with oral transmissions (e.g. legends, quotation of spoken opinions).

D. Editorial markup usually describes uncertain parts (word which cannot be deciphered from the manuscript, damaged places, partially reconstructions, etc.).

Processing Linguistic Vagueness and Uncertainty

For the annotation of linguistic vagueness and uncertainty we use a mixed approach based on collaboration between linguists (one for each language involved) and a computer scientist.

The linguists prepared a list of language-dependent vagueness following Pinkal's classification[16]

- 1. Comparatives, inexact adjectives, for example *'mehr/more,' 'größer/bigger,' 'älter/older'*
- 2. Non-intersectives, for example 'vermeintlich/supposed,' 'so-genannt/so-called'
- 3. Hedges, for example 'ziemlich/quite,' 'einigermaßen/approximately,' 'etwa/about'
- 4. Inexact measures, such as *'4 Tagereisen/4 days' trip,' '10 Fuß /10 feet'*
- 5. Modals (attitudes), for example 'vielleicht/maybe,' 'hoffentlich/hopefully;' and subjunctive verbs
- 6. Lexical quotation markers, such as 'es wurde gesagt /it is said'
- 7. Vague quantifiers, such as 'viele,' 'meistens /mostly'
- 8. Complex quantifiers, for example *'etwa die Hälfte von den 20–30 tausend Soldaten / about a half of the 20–30 thousand soldiers'*
- 9. Numbers
- 10. Range expressions, for example *'Anfang des 18. Jhds./beginning of the 18th century'*
- 12. Unclear person, such as *'der ehemaligen Herzog / the former duke'*
- 13. Unclear time, for example *'in alten Zeiten /in olden times'*

The list of vagueness indicators was created manually and then enriched semi-automatically with elements of synsets (synonyms) extracted from the corresponding language-specific WordNet. The word semi-automatic meant that for each term in the vagueness list, its synset was extracted automatically from the WordNet. Then a human evaluated if the synset elements were part of the vocabulary of the 18th cen-

16 Manfred Pinkal, Logik und Lexikon: *Die Semantik des Unbestimmten* (Berlin/New York: de Gruyter, 1985).

tury or, in a positive case, if the semantics was the same. Finally, we created a vagueness and uncertainty indicator list for each language encoded in XML format.[17]

The annotation of these vagueness markers in the text is facilitated by an automatic morphological annotation, assigning to each token a part-of-speech and a morphological value including, for example comparative degrees for adjectives. In order to ensure uniform annotation across languages, we decided to use the CoNLL-U (Universal Dependencies[18]) format. It was necessary as all the languages involved had a rich morphology and, therefore, inflected/derived/conjugated forms of the items in the vagueness list had to be detected.

Development of a Rich Knowledge Base including Vague and Uncertain Concepts and Relations.

The HerCoRe knowledge base is a type of Fuzzy OWL[19] Ontology trying to model the Ottoman Empire world with its administrative, social, geographical and religious facets.

The following aspects need particular attention:

The modelling of geographical identities and their respective political entities is one such aspect. Political entities (e.g. countries) often tend to share their names with some geographical entities. Political entities retain their names but often change their borders. Thus, we considered as fixed unambiguous individuals those geographical elements still visible today. Historically attested geographical zones which did not exist were modelled as fuzzy concepts.

Political entities were defined as a sum of several historical contexts. Additionally, we introduced the concept of 'historical zone' in order to model concepts such as 'Europe' which from the point of view of the Ottoman Empire, for example began at its borders with Hungary, or 'the Balkans' which for the Ottoman Empire was represented by the Wallachia and Moldavia principalities. In Figure 4, we illustrate the representation of the vague geographical entity of Syrfia which, depending on the historical sources, was situated in three different regions of Europe. Originally we describe three political contexts, each having the same weight. Depending on the additional information (in this case a map), the confidence level can be increased for

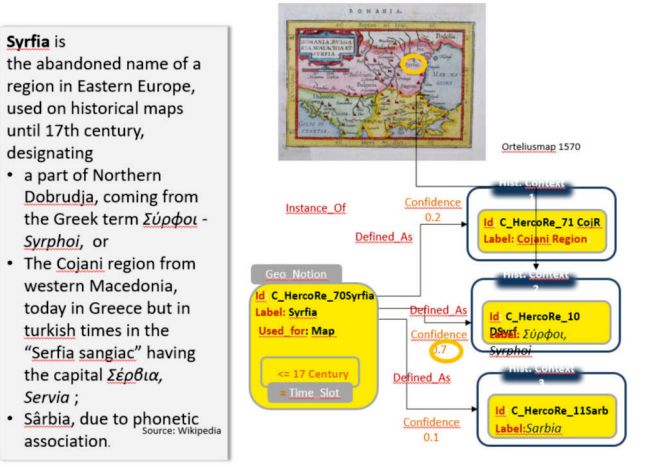
17 Alptug Güney, Walther von Hahn, Cristina Vertan, "Checking reliability of quotations in historical texts – A digital humanities approach", in: Steven Krauwer and Darjy Fiser (Eds.): *Proceedings of Twin Talks 2 and 3, 2020 Understanding and Facilitating Collaboration in Digital Humanities* (2020), https://ceur-ws.org/Vol-2717/.

18 https://universaldependencies.org/format.html.

19 OWL – Web Ontology Language.

one of the hypotheses. In the example in figure 4, the placement on the map near the Black Sea reinforces one of the three hypotheses.

Figure 4: Example of modelling multiple political contexts for one proper name (Syrfia).



Time intervals and geographical positions include fuzzy concepts which allow us to model uncertain dates and coordinates.[20] For the modelling of fuzzy time intervals, we follow the approach described by Métais et al.[21]

A particular challenge is posed by the representation of domain-specific fuzzy concepts. For example, the concept 'place of pilgrimage' is a geographical entity usually containing a monument and being visited by hundreds of people over the years. The complexity of the representation is revealed here by time dependency. A place cannot be an object of pilgrimage from the beginning, but it becomes one (or it loses this qualification) depending on the number of recorded visitors. The cooperation of a human researcher proves unavoidable to determine it: such knowledge cannot be inferred from elsewhere. In collaboration with the researcher in Ottoman studies we defined a threshold for the fuzzy representation of a place of pilgrimage. In

20 Cristina Vertan, "Annotation of vague and uncertain information in historical texts", in: M. Slavcheva et al. (Eds.): *Knowledge, Language, Models* (INCOMA Ltd. Shoumen, Bulgaria, 2020).

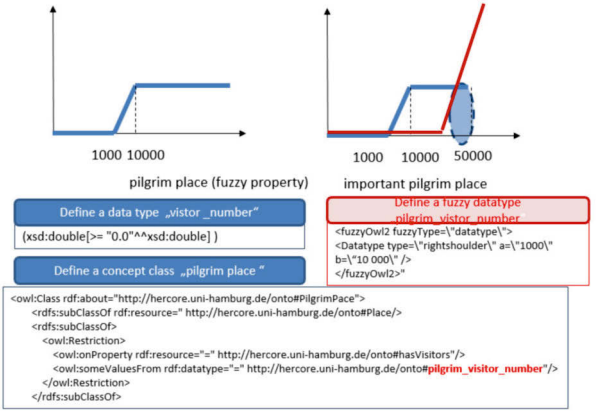
21 Elisabeth Métais, Fatma Ghorbelb, Fayçal Hamdia et al.: "Representing Imprecise Time Intervals in OWL 2", in: *Enterprise Modelling and Information Systems Architectures* Vol. 13, 2018, https://hal.archives-ouvertes.fr/hal-02467393/file/191-Article%20Text-478-1-10-20180322.pdf [accessed: 06.2021].

figure 5, we illustrate this step and show how this is expressed in OWL by using the fuzzy extension.

The ontology contains for the moment more than 500 classes, 250 object properties (relations between classes), and 3,000 individuals.

In order to deal with multilinguality, we attach to each individual the names used in the German translation, the official Ottoman documents and in the Romanian sources.

Figure 5: Fuzzy modelling and representation of the concept 'place of pilgrimage'.



A particular problem in the ontological representation of individuals (concrete representation of classes e.g. Istanbul is an individual of Class 'City') is represented by the multiple—sometimes contradictory—relationships between individuals according to different historical sources.

The model behind OWL is the triple association of <Subject><Predicate><Object> where the <Predicate> is represented by a relationship (e.g. <Constantinople><was_conquered_by><Sultan_Mehmet_II>). This underlying model does not allow the inclusion of a source of truth for this assertion. In our system we overcame this limitation with the following approach:

– the researcher in Ottoman studies identified three major Ottoman works that were used in Cantemir's time (and was also quoted by him) to write the history of the Ottoman Empire. If a certain relation such as 'was_conquered_by' is dependent on historical evidence, we define sub-relations such as 'was_conquered_by_Source1',

'was_conquered_by_Source2', or 'was_conquered_by_Source3'. If there is no historical evidence, we add an additional sub-relation 'was_conquered_by_unclear'. If no sub-relation is defined, we consider the relation as representing a historical fact generally accepted by the research community.

Each individual in the ontology received a unique identifier. This identifier was used for linking individuals with their particular realizations in the corpus.

Representation of Factual Uncertainty

Under 'factual uncertainty' come all paragraphs related to quotations (direct and indirect) from other sources. The quotation style differs dramatically from the modern one. Often paragraphs are quoted but sources are not given. Often quotations are rephrased. There is no bibliographical list. Thus a hermeneutic study is necessary.

Given the fact that DM (Description of Moldavia) practically lacks explicit quotations, the study concentrates on the analysis of HO (History of Ottoman Empire). It follows three directions:

- 1. Identification of the works quoted by Cantemir
- 2. Collection of expressions used by Cantemir when quoting
- 3. Comparison of related facts described by Cantemir against the sources identified in 1).

The study was extremely important as it revealed that not always there was one-to-one correspondence between Cantemir's quotation and the way he reported the event/fact. We identified several cases in which Cantemir had rated a certain event as 'sure' while the available sources reported it contrarily or omitted the fact altogether. On the other hand, there were cases in which Cantemir presented facts, which were reported in other sources, as quite vague or unsure.

This is a very important result which shows us that linguistic vagueness is not the only indicator for assessing the degree of truth for an utterance. The manual annotation, based on the hermeneutic investigation, cannot be avoided. A list of most important works quoted by Cantemir can be found in Vertan's further writings.[22]

Further Work and Discussions

The digital representation and annotation of the corpus was more challenging than expected. We identified the following challenges:

22 Vertan, "Annotation".

- 1. Existent digital versions of some parts of the corpus were not appropriate for our work (see section 3.1)
- 2. The development of the knowledge base took much longer than expected. The collaboration between the researcher in Ottoman studies and the computer scientist was more intensive than expected. The reasons for this are twofold:
  - The most appropriate tool to be used for the development of the ontology is Protégé[23]. The web-based version of the tool is meant to address a non-specialist difficulty and differs from the client version. Thus parallel work on parts of Ontology was impossible
  - The user-friendly fuzzy module for Protégé[24] runs only on an older version of Protégé. On the other hand, only the newest version allows some Description Logics features which we needed. Fuzzy classes and relations have to be written manually in OWL, this operation can be done only by the computer scientist
- 3. The query module for the fuzzy ontology is a programme called reasoner; it is able to use fuzzy logic and realize inferences (judgements). (Our assumption that such a system already existed was partially confirmed.) The reasoner[25] process successfully fuzzy concepts and relations. They, however, fail when processing complex classes modelled with traditional description logics

At the moment of completion of this article, the work concentrates on the development of an appropriate query /interrogation module for the corpus and the presentation of results. The general aim is to present to the user all possible solutions of the query even if not all of them have the same degree of truth (plausibility). The system should only display possible investigation paths; the researcher should use hermeneutic methods for the interpretation.

The development of the ontology turned to be a powerful and new tool for the hermeneutic research in Ottoman studies. At the beginning of the project, we expected that the use of the completed ontology would be of great benefit to the Ottoman studies research. It turned out that the efforts to structure the information—as demanded by the ontology—and the necessity to feed the knowledge base with extensive information triggered the need for further explorations in the

23 https://protege.stanford.edu/.

24 Fernando Bobillo, Miguel Delgado and Juan Gomez-Romero: "Reasoning in Fuzzy OWL 2 with DeLorean, in: Fernando Bobillo, Paulo C. G. Costa, Claudia d'Amato, Nicola Fanizzi, Kathryn B. Laskey, Kenneth J. Laskey, Thomas Lukasiewicz, Matthias Nickles and Michael Pool (Eds.): *Uncertainity Reasoning for the Semantic Web II* (Berlin/Heidelberg: Springer Verlag, 2013).

25 Fernando Bobillo and Umberto Straccia, "Fuzzy Ontology Representation using OWL 2", ht tp://arxiv.org/pdf/1009.3391.pdf [accessed: 15.02.2016].

archives. A PhD in Ottoman studies on Dimitrie Cantemir's works will be completed soon as a direct result of this project.

In parallel, we established a network of researchers working on applying fuzzy reasoning on real data, in particular humanities. It turned out that many of the reported solutions were tested on small, less complex data and that the systems indeed had limitations on how to handle more complex corpora as the ones presented here. We expect that parts of the HerCore corpus as well as the ontology will be used in the future for system improvements.

The annotation tool, the ontology and a limited part of the annotated corpus will be publicly available under CC BY licence.

Acknowledgments

The author thanks all participants in and contributors to the HerCoRe project: Alptug Güney (University of Samsun, formerly of University of Hamburg), Walther von Hahn (University of Hamburg), Ioana Costa (University of Bucharest), Yavuz Köse (University of Vienna), Rafael Quirros Marin and Miguel Pedregosa (University of Granada, formerly of University of Hamburg).

Bibliography

Bobillo, Fernando and Umberto Straccia, "Fuzzy Ontology Representation using OWL 2", http://arxiv.org/pdf/1009.3391.pdf [accessed: 15.02.2016].

Bobillo, Fernando, Miguel Delgado and Juan Gomez-Romero, "Reasoning in Fuzzy OWL 2 with DeLorean, In *Uncertainity Reasoning for the Semantic Web II* edited by Fernando Bobillo, Paulo C. G. Costa, Claudia d'Amato, Nicola Fanizzi, Kathryn B. Laskey, Kenneth J. Laskey, Thomas Lukasiewicz, Matthias Nickles and Michael Pool, (Berlin/Heidelberg: Springer Verlag 2013).

Cantemir, Dimitrie, *Beschreibung der Moldau*, Faksimiledruck der Originalausgabe von 1771 (Frankfurt und Leipzig, 1771).

Cantemir, Dimitrie, *Geschichte des osmanischen Reichs nach seinem Anwachse und Abnehmen*, (Hamburg: Herold, 1745).

Costa, Ioana and Florentina Nicolae (eds.): *Descrierea stării Moldaviei* (Bucureşti: Academia Română-Fundaţia Naţională pentru Ştiinţă şi Artă, 2017).

Gordon, Octavian, Florentina Nicolae, Monica Vasileanu and Ioana Costa (eds.): *Cantemir, Dimitrie, Istoria măririi şi decăderii Curţii othmane*, 2 volumes, (Bucureşti, Academia Română-Fundaţia Naţională pentru Ştiinţă şi Artă, 2015).

Güney, Alptug, Walter von Hahn and Cristina Vertan: "Checking reliability of quotations in historical texts – A digital humanities approach", In *Proceedings of Twin Talks 2 and 3, 2020 Understanding and Facilitating Collaboration in Digital Humanities*

edited by Steven Krauwer and Darjy Fiser (2020), https://ceur-ws.org/Vol-2717/.

Métais, Elisabeth, Fatma Ghorbel, Fayçal Hamdi, Nebrasse Ellouze, Noura Herradi and Assia Soukane, "Representing Imprecise Time Intervals in OWL 2", In *Enterprise Modelling and Information Systems Architectures* Vol. 13, 2018, https://hal.archives-ouvertes.fr/hal-02467393/file/191-Article%20Text-478-1-10-20180322.pdf [accessed: 06.2021].

Pinkal, Manfred, *Logik und Lexikon: Die Semantik des Unbestimmten* (Berlin/New York: de Gruyter, 1985).

Vertan, Cristina, "Annotation of vague and uncertain information in historical texts", In *Knowledge, Language, Models* edited by M. Slavcheva et al., (INCOMA Ltd. Shoumen, Bulgaria, 2020).

von Hahn, Walther, „Vagheit bei der Verwendung von Fachsprachen", In: *Fachsprachen. Ein Internationales Handbuch zur Fachsprachenforschung und Terminologiewissenschaft* edited by Lothar Hoffmann, Hartwig Kalverkämper and Herbert Ernst Wiegand, (Berlin 1998), 383–390.

#HETEROGENEITY

#Heterogeneity describes "(t)he quality or condition of being heterogeneous", either as a "(d)ifference or diversity in kind from other things" or in the sense of a "(c)omposition from diverse elements or parts". * There is no specific derivation yet addressing this phenomenon in computer science.

It contrasts with the role of the term in digital humanities, more so in the context of the mixed methods projects. There is the challenge of "the heterogeneity of data demand and data supply" in computer science that might lead to ignoring a #Corpus that falls short of the demand (ArchiMediaL), thus preventing innovative approaches.

On the other hand, #Heterogeneity is in fact one of the parameters that make the cooperation between humanities and computer science such a fascinating endeavour. Here, "(d)ata is heterogeneous with respect to encoding formats, languages, scripts, mode (picture, text, video), and historical language variants", thus creating a challenge for conceptualization and application (HerCoRe).

* "Heterogeneity, n.". in: *Oxford English Dictionary* (OED), first published 1898; most recently modified version published online March 2022, https://www.oed.com/ [accessed: 20.05.2022].

Authors

Melanie Andresen is a Postdoc researcher at the Institute for Natural Language Processing at the University of Stuttgart. After studying German Language and Literature (B.A.) and German Linguistics (M.A.) at Universität Hamburg, she specialized in corpus linguistics and digital humanities in her dissertation on data-driven corpus analyses. She has worked in several interdisciplinary DH projects that explored the benefit of methods from corpus and computational linguistics for disciplines such as literary studies, cultural anthropology, nursing science, and philosophy. Her current fields of research include German academic discourse and computational literary studies with a focus on drama analysis.

Ralph Barczok studied Catholic Theology, Languages and Cultures of the Christian Orient in Tübingen and Paris and received his PhD in History from the University of Konstanz in 2018 with a thesis on monasticism in medieval northern Iraq. From 2013 to 2017 he worked at the University of Konstanz and from 2017 to 2022 at the Goethe University Frankfurt as a research assistant on various projects. His research focuses on the Christianities of the Middle East, the form and function of their ecclesiastical elites and the monastic movements there.

Timo Baumann is a professor in Artificial intelligence and Natural Language Processing at Ostbayerische Technische Hochschule Regensburg focusing on spoken language interaction. Timo studied computer science and phonetics in Hamburg, Geneva and Granada and received his PhD for work on incremental spoken dialogue processing from Bielefeld University. Timo held posts as researcher at Potsdam, Bielefeld and Hamburg universities and worked as a systems scientist at Carnegie Mellon University during which time he co-headed the Rhythmicalizer research group on the computer-aided classification of poetic prosodies. Prior to his current position, Timo was the interim professor in Machine Learning at Hamburg University.

Ulrik Brandes is professor for social networks at ETH Zürich. His background is in computer science (diploma 1994 RWTH Aachen, doctorate 1999 and habilitation 2002 University of Konstanz). He is vice-president of the International Network for Social Network Analysis (INSNA), coordinating editor of Network Science, and on the editorial boards of Social Networks, Journal of Mathematical Sociology, Journal of Graph Algorithms and Applications, and Computer Science Review. He is a co-author of the visone software for network analysis and the GraphML data format. Major grants include a DFG Koselleck Project (2011–2017) and an ERC Synergy Project (2013–2019). His interests are in network analysis and visualization broadly, with applications to social networks in particular. Currently, his focus is on collective behaviour in association football (soccer).

Victor de Boer is an Associate Professor of User-Centric Data Science at Vrije Universiteit Amsterdam and he is Senior Research Fellow at the Netherlands Institute for Sound and Vision. His research focuses on data integration, semantic data enrichment and knowledge sharing using Linked Data technologies in various domains. These domains include Cultural Heritage, Digital Humanities and ICT for Development where he collaborates with domain experts in interdisciplinary teams. Victor has been involved (and is currently involved) in various National and European research projects, focusing on user-centric data integration and semantic enrichment. He is one of the co-directors of the Cultural AI lab.

Max Franke received his Bachelor and Master of Science in computer science from the University of Stuttgart, Germany. He is currently working towards a doctorate degree at the department for visualization and interactive systems of the University of Stuttgart. His research focusses on visually and algorithmically supporting analyses performed by Digital Humanities scholars, especially in research domains where (geo)spatial and temporal data are of interest.

Christoph Finkensiep is a doctoral researcher at the Digital and Cognitive Musicology Lab of the École Polytechnique Fédérale de Lausanne. He obtained his Bachelor's degree in Computer Science at the University of Paderborn (2014) and completed his Master's degree in Cognitive Science at the University of Osnabrück with a thesis entitled "A Formal Model of Voice Leading" (2017). His dissertation "The Structure of Free Polyphony" (2022) proposes a computational model of tonal structure on the level of notes. Research interests include music cognition, probabilistic modelling and machine learning, artificial intelligence, as well as philosophy of mind and philosophy of science.

Jan van Gemert received a PhD degree from the University of Amsterdam in 2010. There he was a post-doctoral fellow as well as at École Normale Supérieure in Paris.

Currently he leads the Computer Vision lab at Delft University of Technology. He teaches the Deep learning and Computer Vision MSc courses. His research focuses on visual inductive priors for deep learning for automatic image and video understanding. He published over 100 peer-reviewed papers with more than 6,000 citations.

Anastasia Glawion is a postdoctoral researcher at the Department of German Studies – Digital Literary Studies at the TU Darmstadt. She studied Sociology and Social Anthropology in St. Petersburg State University (2011), has a Master's Degree in European Cultural Studies from Constance University (2014) and finished her PhD in 2021 (TU Darmstadt). Her interests are in digital literary studies, network analysis and its application to literary and cultural studies, the empirical studies of literature, emotion, and reader response.

Robert Hammel obtained a master's degree in Slavic and Romance (Italian) philology at the University of Frankfurt am Main in 1990 (M. A. thesis: Contrastive Studies on Russian Verbs of Motion within the Framework of Functional Grammar). From 1992–1997 he was research assistant at the Department of Slavic Studies of the University of Göttingen where he completed his PhD thesis on the development of Russian present and past tense inflection in 1996. From 1998–2004 he was postdoctoral research Assistant at the Department of Slavic Studies of Humboldt University of Berlin. Since 2007 he is lecturer in West Slavic linguistics at the Department of Slavic and Hungarian Studies of Humboldt University of Berlin.

Carola Hein is Professor and Head, Chair History of Architecture and Urban Planning at Delft University of Technology. She has published widely in the field of architectural, urban and planning history and has tied historical analysis to contemporary development. Among other major grants, she received a Guggenheim Fellowship to pursue research on The Global Architecture of Oil and an Alexander von Humboldt fellowship to investigate large-scale urban transformation in Hamburg in international context between 1842 and 2008. Her current research interests include the transmission of architectural and urban ideas, focusing specifically on port cities and the global architecture of oil. She has curated Oildam: Rotterdam in the oil era 1862–2016 at Museum Rotterdam. She serves as IPHS Editor for Planning Perspectives and as Asia book review editor for Journal of Urban History. Her books include: *The Routledge Planning History Handbook* (2017), *Uzō Nishiyama, Reflections on Urban, Regional and National Space* (2017), *History, Urbanism, Resilience, Proceedings of the 2016 IPHS conference* (2016), *Port Cities: Dynamic Landscapes and Global Networks* (2011), *Brussels: Perspectives on a European Capital* (2007), *European Brussels. Whose capital? Whose city?* (2006), *The Capital of Europe. Architecture and Urban Planning for the European Union* (2004), *Rebuilding Urban Japan after 1945* (2003), and *Cities, Autonomy*

and Decentralisation in Japan. (2006), *Hauptstadt Berlin 1957–58* (1991). She has also published numerous articles in peer-reviewed journals, books, and magazines.

Paul Heinicker is a design researcher investigating discursive design concepts with a focus on the culture and politics of diagrams and data visualisations. His practice covers image-led research as well as written analyses. He finished his PhD at the University of Potsdam at the Institute for Media and Art and has an interdisciplinary background in multimedia technologies (B.Eng.) and interface design (M.A.).

Katharina Herget is a research assistant at the Department of German Studies – Digital Literary Studies headed by Prof. Dr. Thomas Weitin at the Technical University of Darmstadt. She studied German Literature and Sociology (B.A.) and holds a Master of Arts in German Literature from the University of Constance. Her research interests include digital approaches to literary history, such as topic modeling and the combination of stylometric data and network analysis, as well as novella research and 19th century literature.

Hussein Hussein studied Electrical Engineering, Computer Technology and Automatic Control from 1992 to 2000 at Tishreen University, Latakia, Syria, and then joined Dresden University of Technology (TUD) from 2004 to 2007 as a master student in acoustic and speech communication. He continued towards PhD studies at TUD which he completed in 2013. Dr. Hussein worked as research assistant at different universities (Berliner Hochschule für Technik, TU Chemnitz, Freie Universität Berlin) as well as in the industry as software developer (Linguwerk GmbH) and development engineer and project manager (ICE Gateway GmbH). He presently works as project manager at atene KOM GmbH.

Seyran Khademi is an Assistant Professor at the faculty of Architecture and the Built Environment (ABE) and the co-director of AiDAPT lab (AI for Design, Analysis, and Optimization in Architecture and the Built Environment). She is working as an interdisciplinary researcher between Computer Vision lab and Architecture Department at ABE. Her research interest lies at the intersection of Data, Computer Vision and Deep Learning in the context of man-made imagery including illustrations and visual data for Architectural Design. In 2020 she was honored to be the research in residence fellow at the Royal Library of the Netherlands working on visual recognition for children's book collection. In 2017 she was appointed as a postdoctoral researcher at Computer vision lab working on the ArchiMediaL project, regarding the automatic detection of buildings and architectural elements in visual data focusing on Computer Vision and Deep Learning methods for archival data and street-view imagery. Seyran received her Ph.D. in signal processing and optimization in 2015 from TU Delft, followed by postdoctoral research on Intelligent Audio and Speech

algorithms. She received her MSc. degree in Signal Processing from the Chalmers University of Technology in Gothenburg, Sweden, in 2010 and her BSc degree in telecommunications from the University of Tabriz in Iran.

Janna Kienbaum studied Italian philology and cultural studies at the University of Potsdam and Humboldt University in Berlin. Her research focuses on diagrammatics, data visualizations and digital collections of museums. In her PhD thesis she investigates the web presentation of museum art collections as a diagrammatic representation system. From 2017–2020, she was a research assistant in the mixed-methods project *analyzing networked climate images (anci)*. For the exhibition *Nach der Natur* at the Humboldt Forum (July 2021), she supervised the exhibition segments "Googled+" and "Klimazukünfte". Since 2021, she has been working as a research assistant in the project "FDNext" for research data management at the University of Potsdam.

Steffen Koch is a research associate at the Institute for Visualization and Interactive Systems, University of Stuttgart, Germany, where he received his doctorate in computer science in 2012. His research interests comprise visualization in general, with foci on visual analytics for text and documents, visualization in the digital humanities, as well as interactive visualization support for data mining and machine learning.

Benjamin Krautter studied German language and literature and political science in Stuttgart and Seoul. Currently, he is a PhD student at the Department of German Studies at the University of Heidelberg and a member of the QuaDramA / Q:TRACK project at the University of Cologne. Among other things, he is working on the operationalisation of literary concepts for quantitative drama analysis. Doing so, he focuses on how to meaningfully combine quantitative and qualitative methods for the analysis and interpretation of literary texts.

Beate Löffler received an engineering degree in Architecture in Potsdam and studied History and History of Art in Dresden afterwards. She was a long-serving employee and project manager of an ethnological digitalization project, which resulted in interests in both the epistemics of image databases and the (trans)cultural exchange of knowledge. She gained her PhD with a book about the acculturation of Christian church architecture in modern Japan (2009) and her habilitation in history and theory of architecture and construction with an analysis of Japan-related architectural discourses in Europe in the late 19th century (2020). Other fields of interest are the religious topography of the contemporary urban and the methodologies of interdisciplinary study of space and architecture. Beate Löffler researches and teaches at the TU Dortmund University.

Tino Mager is Assistant Professor of the History and Theory of Architecture and Urbanism at the University of Groningen, and President of ICOMOS Germany. Previously, he worked at the Faculty of Architecture and Built Environment at Delft University of Technology, was a fellow of the Leibniz Association and the University of Queensland. He studied media technology in Leipzig and art history and communication science in Berlin, Barcelona and Tokyo. He is Secretary General of the ICOMOS International Scientific Committee on Water and Heritage and published widely on cultural heritage. His books include: *Schillernde Unschärfe – Der Begriff der Authentizität im architektonischen Erbe* (De Gruyter 2016), *Architecture RePerformed: The Politics of Reconstruction* (Routledge 2015), *Water Heritage: Global Perspectives for Sustainable Development* (BOCH 2020), *Rational visions – production of space in the GDR* (Bauhaus University Press 2019), *BetonSalon – New Positions on Late Modern Architecture* (Neofelis 2017) and *Church buildings and their future. Restoration – conversion – adaptive reuse* (Wüstenrot Foundation 2017).

Roland Meyer is professor of West Slavic linguistics at Humboldt-Universität zu Berlin (since 2012). He holds master's degrees in Slavic and computational linguistics from the University of Tübingen, where he also completed his PhD on the syntax of questions in Russian, Polish and Czech (2002). From 2003 to 2011, he was assistant professor ('Akademischer Rat') of Slavic linguistics at the University of Regensburg, where he finished his habilitation on the history of null subjects in Russian, Polish and Czech (2011) and acted as stand-in professor of West-Slavic linguistics in 2012. Roland Meyer has headed several research projects with a strong computational or corpus linguistic part on case in Slavic, register in Slavic, on bias in Slavic questions, and on Slavic language history.

Burkhard Meyer-Sickendiek co-headed the Rhythmicalizer research group funded by Volkswagen Foundation. He earned his doctorate at Tübingen University with a study on the "Aesthetics of Epigonality" and habilitated on "literary sarcasm in German-Jewish modernity" at LMU Munich. In 2008, Meyer-Sickendiek joined FU Berlin as a guest professor within the "Languages of Emotion" cluster of excellence. We works in the broader subject area of so-called "affect poetics" and published several monographs and anthologies: For example, on "lyrical intuition", the poetology of "rumination" or the theater history of "tenderness". His current study on "Hörlyrik" examines portals for audio poems.

Markus Neuwirth is a Professor of Music Analysis at the Anton Bruckner University Linz (since 2020). Previously he held postdoctoral positions at the Digital and Cognitive Musicology Lab of the École polytechnique fédérale de Lausanne (EPFL) and the University of Leuven, where he obtained his PhD in musicology in 2013. He is co-editor of the journal *Music Theory and Analysis*, as well as one of the main editors

of the *GMTH Proceedings*. In addition, he has been co-editor (with Pieter Bergé) of the volume *What is a Cadence? Theoretical and Analytical Perspectives on Cadences in the Classical Repertoire* (Leuven University Press, 2015), which received the Outstanding Multi-Author Collection Award 2018 from the Society for Music Theory. Neuwirth is the co-author (with Felix Diergarten) of a musical *Formenlehre* that has been published with Laaber in 2019.

Thomas Nocke is a senior researcher at the Potsdam Institute for Climate Impact Research. He studied Computer Science / Computer Graphics at the University of Rostock and did his phd there on the topic of Climate Data Visualization. Since he is with the Potsdam Institute, he investigated visual analytics of large climate data and visual communication of climate science knowledge. He is interested in interdisciplinary approaches, including digital humanities and digital climate services.

Janis Pagel is a PhD student at the Institute for Natural Language Processing at the University of Stuttgart and research associate at the Department for Digital Humanities at the University of Cologne. He studied German studies and linguistics in Bochum, and computational linguistics in Stuttgart and Amsterdam. His research focuses on the application of computational linguistic methods to concepts from literary studies and coreference resolution on literary texts.

Simon Päpcke is a doctoral researcher at the Social Networks Lab at ETH Zurich. He studied mathematics at Kiel University and ETH Zurich (MSc ETH 2016) and worked in a data science team at a Swiss insurance company. In 2022, he defended his dissertation at ETH Zurich. His research interests are in network analysis, natural language processing, and computational text and corpus analysis.

Marcus Pöckelmann studied computer science at the Martin Luther University Halle-Wittenberg (Master 2013) and has been a member of the research group Molitor/Ritter since 2013. Within several interdisciplinary research projects, he develops web-based applications for the investigation of intertextuality together with colleagues from different disciplines of the humanities. These include the working environments LERA for the analysis of complex text variants for scholarly editions, and Paraphrasis for the retrieval and evaluation of paraphrased text passages in the ancient Greek literature.

Andrew Prescott is Professor of Digital Humanities in the School of Critical Studies, University of Glasgow. He trained as a medieval historian and from 1979–2000 was a Curator in the Department of Manuscripts of the British Library, where he was the principal curatorial contact for Kevin Kiernan's *Electronic Beowulf*. Andrew was from 2012–2019 Theme Leader Fellow for the AHRC strategic theme of *Digital Trans-*

formations. He has also worked in libraries, archives and digital humanities units at the University of Sheffield, King's College London and the University of Wales Lampeter. Publications include *English Historical Documents* (1988), *Towards a Digital Library* (1998), *The British Inheritance* (2000) and *Communities, Archives and New Collaborative Practices* (2020), as well as numerous articles on digital humanities, the history of libraries and archives, and medieval history.

Nils Reiter studied computational linguistics and computer science at Saarland University. He did his PhD in a collaboration project between classical Indology and computational linguistics at Heidelberg University (CRC "ritual dynamics") and then worked at Stuttgart University as a scientific coordinator and investigator in the Centre for Reflected Text Analytics (CRETA). Since October 2021 he is Professor for Digital Humanities and Computational Linguistics at the University of Cologne, and head of the Data Center for the Humanities which provides research data management services and consulting to the faculty. His research interests are related to operationalization, particularly with respect to questions and concepts from literary studies.

Martin Rohrmeier studied musicology, philosophy, and mathematics at the University of Bonn and earned an MPhil and PhD in musicology at the University of Cambridge/UK. Having been a postdoctoral researcher at Microsoft Research, FU Berlin and the Massachusetts Institute of Technology, he joined TU Dresden as Open-Topic-Professor for music cognition in 2014. Since 2017 he is Professor for Digital Musicology at the École Polytechnique Féderale de Lausanne, where he directs the Digital and Cognitive Musicology Lab (DCML). The central research projects lie at the intersection of music theory, cognition, and computation and have been funded by the Volkswagen Foundation, the SNF, and the ERC Starting Grant. Main areas of research are digital musicology, formal music theory and analysis, music psychology and cognition, as well as philosophy of language and music.

Birgit Schneider is professor for Knowledge Cultures and Media Environments in the Department of European Media Studies at the University of Potsdam, Germany. She studied art and media studies as well as media art and philosophy in Karlsruhe, London and Berlin. After initially working as a graphic designer, she worked from 2000 to 2007 at the research department "The Technical Image" at the Humboldt University in Berlin, where she received her doctorate. Since 2009, she has been researching in the context of fellowships at the European Media Studies Department of the University of Potsdam as well as in Munich, Weimar and Cambridge, UK. In 2010 she represented the Chair of History and Theory of Cultural Techniques at the Bauhaus University Weimar. Her current research focuses are images and percep-

tions of nature, ecology and climate change, diagrams, data graphics and maps as well as images of ecology. She is head of the mixed-methods project "analysing networked climate images", co-speaker of the "Network Digital Humanities" of the University of Potsdam and a member of the research group "Sensing. On the knowledge of sensitive media". A selection of publications: "The Technical Image" (Cambridge 2015) and Image Politics of Climate Change" (Bielefeld 2014) and the German monographs "Textiles Prozessieren" (Berlin, 2007) and "Klimabilder" (Berlin 2018).

Ronald Siebes is a senior researcher at the User Centric Data Science group in the department Computer Science at the VU Amsterdam. He applies Linked Data research and Machine Learning in various national and European projects, ranging from Social Sciences, Humanities to IoT. Within ArchiMediaL he developed a crowd sourcing platform where participants via a web application annotate and compare historical street view images with current street view images. This data resulted in a valuable bench mark dataset for the Visual Machine Learning research community. Ronald received his Ph.D. in Artificial Intelligence where he worked on distributed reasoning algorithms using Peer-to-Peer technology.

Vera Szöllösi-Brenig has been programme director at the Volkswagen Foundation since 1999. She studied German and French Literature as well as Linguistics in Munich, passed a traineeship at a public radio station and worked for a decade as political journalist (anchor) at the Deutschlandfunk. In this time, she wrote her thesis in French Literature (Nouveau Roman). At the Volkswagen Foundation, she has been in charge of a broad range of funding programmes, e.g. "Documentation of Endangered Languages", "Key Issues in the Humanities" and the call "Mixed Methods – Support for Projects Combining and Synergizing Qualitative-Hermeneutical and Digital Approaches". At present, her range of tasks include the coordination of the Open Science activities at the Volkswagen Foundation.

Aleksej Tikhonov is a linguist of the Department of Slavonic and Hungarian Studies at the Humboldt University of Berlin. He completed his Ph.D. on the linguistic author identification of Rixdorf manuscripts in 2020. Currently he is working as PostDoc of the UK-German collaborative project "The History of Pronominal Subjects in the Languages of Northern Europe" between the Humboldt University of Berlin (head: Roland Meyer) and the University of Oxford (head: David Willis), and in the Multilingual Handwritten Text Recognition Project at the University of Freiburg (head: Achim Rabus). His focus languages are Russian, Czech, German, Polish, Ukrainian, and Yiddish.

Cristina Vertan is senior researcher at the Digital Humanities Research Group of the Berlin Academy of Sciences. She holds a Ph.D. in Computer Science from the

University of Bucharest, Romania and a Master in Statistics from the Free University of Brussels, Belgium. Following a Humboldt grant, she conducted research for over twenty years in Natural Language Processing and digital Humanities at the University of Hamburg, Germany. Her field of expertise are multilingual applications, data modelling and computational analysis of historical texts. In these fields she co-authored more than hundred publications, leaded national and international funded research projects dealing with various less-resourced and historical languages, among them the project HerCoRe (Hermeneutic and Computer-based Analysis of Reliability, Consistency and Vagueness in historical texts) funded by the Volkswagen Foundation.

Thomas Weitin is professor for digital philology at TU Darmstadt. His background is in experimental literary analysis with an emphasis on law and literature (magister artium 1997 University of Hamburg, doctorate 2002 Humboldt University Berlin and habilitation 2008 University of Münster). He is Senior Editor of the Open Library of the Humanities and editor of the series Digitale Literaturwissenschaft with Springer Nature and Recht und Literatur with Nomos. He is founder and chair of the Darmstadt Litlab for cognitive reception analysis. He was Humboldt Fellow at the Johns Hopkins University Baltimore, Max Planck-Fellow and Senior Fellow at the International Center for Cultural Studies in Vienna. He was visiting professor at the University of California, Berkeley and at the Tongji University in Shanghai. Major Grants include two DFG projects and a Volkswagen „Schlüsselthemen der Geisteswissenschaften". Latest book: Digitale Literaturwissenschaft. Eine Versuchsreihe mit sieben Experimenten. Springer Nature 2022.

Dorothea Weltecke has held the Chair of Medieval History at Humboldt-Universität zu Berlin since 2021. Before she held chairs for the History of Religions at the Universität Konstanz (2007–2017) and for Medieval History at Goethe-Universität Frankfurt a. Main (2017–2021). Dorothea Weltecke studies the inter- and intra-religious dynamics in the history of religions in Europe and the Middle East and particular focuses on the centuries between 500 and 1500. Since her PhD thesis on the Syriac orthodox Patriarch Michael the Great (1126–1199) research on Eastern Christianity has been of interest to her.

Marcus Willand studied (2002–08) linguistics and literature, psychology and sociology in Darmstadt, Berlin and Turku (Finland). PhD (HU-Berlin, 2009–2014) on 'Reader models and reader theories', former member of the PhD-Net: 'The Knowledge of Literature' and scholarship holder of the doctoral funding of the Studienstiftung des deutschen Volkes (German National Academic Foundation). From 2013 to 2020 research assistant of A. Albrecht in Stuttgart and Heidelberg. 2014 to 2018 editor of Scientia Poetica. Former project manager (with Nils Reiter) of 'QuaDramA'

(Volkswagen Foundation) and 'Q:TRACK' (Priority Program 'Computational Literary Studies', DFG).

Eva Wöckener-Gade holds a PhD in in Classical Philology. Her research interests focus on the study of ancient Greek literature of the classical period and its reception throughout antiquity and beyond. Methodically, she is committed to working in (interdisciplinary) teams and applying new digital methods to rather old questions; she has so far been able to practise this in the projects eXChange and Digital Plato (both at the University of Leipzig) and recently as a team member in the project 'Etymologika' at the University of Hamburg (https://www.etymologika.uni-hamburg.de/).

[transcript]

PUBLISHING.
KNOWLEDGE. TOGETHER.

transcript publishing stands for a multilingual transdisciplinary programme in the social sciences and humanities. Showcasing the latest academic research in various fields and providing cutting-edge diagnoses on current affairs and future perspectives, we pride ourselves in the promotion of modern educational media beyond traditional print and e-publishing. We facilitate digital and open publication formats that can be tailored to the specific needs of our publication partners.

OUR SERVICES INCLUDE

- partnership-based publishing models
- Open Access publishing
- innovative digital formats: HTML, Living Handbooks, and more
- sustainable digital publishing with XML
- digital educational media
- diverse social media linking of all our publications

Visit us online: www.transcript-publishing.com

Find our latest catalogue at www.transcript-publishing.com/newbookspdf